A Systems Biology Approach To

Musculoskeletal Tissue Engineering:

Transcriptomic And Proteomic Analysis Of

Cartilage And Tendon Cells

Thesis submitted in accordance with the requirements of the

University of Liverpool for the degree of Doctor of Philosophy

by

Alan James Mueller

May 2015

Abstract

Disorders of cartilage and tendon account for a high incidence of disability and are highly prevalent co-morbidities within the ageing population; therefore, musculoskeletal disorders represent a major public health policy issue. Despite considerable efforts to characterise biochemical and biomechanical cues that promote a stable differentiated cartilage or tendon phenotype *in vitro* the benchmarks by which progress is measured are limited. Common regenerative interventions, such as autologous cartilage implantation, have a required period of monolayer expansion that induces a loss of the functional phenotype, termed dedifferentiation. Dedifferentiation has no definitive mechanism yet is widely described in both regenerative and degenerative contexts; in addition to stem cell transplantation and cell-seeding in three-dimensional scaffolds, dedifferentiation represents the third approach to the development of regenerative mechanisms for mammalian tissue repair.

Cartilage and tendon show a number of common features in structure, develop, disease, and repair. The extracellular matrix is a dynamic and complex structure that confers the functional mechanical properties of cartilage and tendon. Dysregulation of production and degradation are critical to the pathophysiology of musculoskeletal disorders, therefore, reparative interventions require a stable, functional phenotype from the outset. Cartilage and tendon demonstrate a commonality in terms of function defining structure both being sparsely cellular with a preponderance of collagenous matrix. Parity of functionality with the pre-injury state after healing is rarely achieved for cartilage and tendon. Cartilage and

tendon also share common embryological origins. Common mesenchymal progenitor cells differentiate into many musculoskeletal tissues with diverse functions. Specialist sub-populations of tendon and cartilage progenitors enable formation of transitional zones between these developing tissues. The development of musculoskeletal structures does not occur in isolation, however, cartilage and tendon have not previously been considered together in a systems context. An integrated understanding of the differentiation of these tissues should inform regenerative therapies and tissue engineering strategies.

Systems biology is paradigm shift in scientific thinking where traditional reductionist strategies to complex biological problems have been superseded by a holistic philosophy seeking to understand the emergent behavior of a system by the integrative and predictive modeling of all elements of that system. Whole transcriptome and proteome profiling studies are used to collect quantitative data about a system, which may then be exploited by systems biology methodologies including the analysis of gene and protein networks. Gene-gene co-expression relationships, which are core regulatory mechanisms in biology, are often not part of a comprehensive gene expression analysis. Many biological networks are sparse and have a scale-free topology, which generally indicates that the majority of genes have very few connections, whilst certain key regulators, or 'hubs', are highly interconnected. Co-expression networks may be used to define regulatory sub-networks and 'hubs' that have phenotypic associations. This approach allows all quantitative data to be used and makes no *a priori* assumptions about relationships in the system and, therefore, can facilitate the exploration of emergent behavior in the system and the generation of novel hypotheses.

The ultimate goal of tissue engineering is the replacement of lost or damaged cells, and *in vitro*, to develop biomimetic (organotypic) structures to serve as experimental models. Tissues, and the strategies to functionally replicate them *ex vivo*, are complex and require an integrated, multi-disciplinary approach. Systems biology approaches, using data arising from multiple-levels of the biological hierarchy, can facilitate the development of predictive models for bioengineered tissue. The iterative refinement, quantification, and perturbation of these models may expedite the translation of well-validated organotypic systems, through legal regulatory frameworks, into regenerative strategies for musculoskeletal disorders in humans.

In this thesis the systems under consideration are the major cell populations of cartilage and tendon (chondrocytes and tenocytes, respectively). They are described in three environmental conditions: native tissue, monolayer (two-dimensional), or three-dimensional models. There has been no systematic investigate of the global gene and protein profiles of cartilage and tendon in their native state relative to monolayer or three-dimensional cultures. There is no clear mechanistic description of the impact of *in vitro* environmental perturbations on the system or indeed the adequacy of these models as proxies for cartilage and tendon.

A discovery approach using transcriptomic and proteomic profiling is undertaken to define a robust and consistent gene and protein profile for each condition. Differentially expressed elements are functionally annotated and pathway topology approaches employed to predict major signalling pathways associated with the observed phenotype. This study defines dedifferentiated chondrocytes and tenocytes in monolayer culture as expressing markers of musculoskeletal

development, including scleraxis (*Scx*) and Mohawk (*Mkx*). Furthermore, there is reproducible synthetic profile convergence in monolayer culture between cartilage and tendon cells. Standard three-dimensional culture systems for chondrocyte and tenocytes fail to replicate the gene expression profile of cartilage and tendon. The PI-3K/Akt signaling pathway is predicted to be the predominant canonical pathway associated with de- and re-differentiation *in vitro*.

Using novel, and publically available, transcriptomic data sets a meta-analysis of microarray gene expression profiles is performed using weighted gene co-expression network analysis. This is employed for transcriptome network decomposition to isolate highly correlated and interconnected gene-sets (modules) from gene expression profiles of cartilage and tendon cells in different environmental conditions. Sub-networks strongly associated with de- and re-differentiation phenotypes are defined. Comparison of global transcriptome network architecture was performed to define the conservation of network modules between a model species (rat) and human data. In addition to the annotation of an osteoarthritis-associated module in the rat a class-prediction analysis defined a minimal gene signature for the prediction of three-dimensional cultures from standard monolayer culture. Finally, proteomic and transcriptomic data sets are integrated by defining common upstream regulators (TGFB and PDGF BB) and unified mechanistic networks are generated for de- and re-differentiation.

The studies collected in this thesis contribute to a wider understanding of cartilage and tendon tissue engineering and organotypic culture development. A clear mechanistic understanding of the regulatory networks controlling differentiation of cartilage and tendon progenitor cells is required in order to develop improved *in*

vitro models and bio-engineered tissue that are physiologically relevant. The findings presented here provide practical outputs and testable hypotheses to drive future evidence-based research in organotypic culture development for musculoskeletal tissues.

Declaration

The work presented in this thesis comprises the author's original contribution towards the PhD except where this is indicated in the Preface.

All other material used is acknowledged in the text

This thesis is fewer than 100,000 words in length exclusive of bibliographies and appendices.

Preface

This thesis is the author's original work except for the following contributions:

Chapter 2:

BBSRC Research Experience Placement (2012): Ms Alex Tanner – performed all qPCR experiments. Project design, primer validation, sample preparation and data analysis was undertaken by the author.

Histology sections: Ms Valerie Tilston – preparation and staining of histological sections.

Microarrays – The Genome Centre, Queen Mary University, London.

Chapter 3:

BBSRC Research Experience Placement (2013): Mr Andrew Rich – preliminary validation of siRNA transfection protocol upon which the author's further work was based. Project design, primer validation, sample preparation and data analysis was undertaken by the author.

Microarrays – Provided as a service by Tepnel Pharma Services, Hologic, Manchester

Chapter 6

Proteomics: Dr. Deborah Simpson – sample standardization, proteolytic digest, liquid chromatography and data acquisition by tandem mass spectrometry undertaken as a service at the University of Liverpool.

Acknowledgements

To the individuals listed here I extend immense gratitude for your expertise and willingness to share your knowledge; your patience and support; your good humour and company; your hospitality and your horses.

Family

Michael, Mum&Dad, LaddieGaga

Academic

Dr. Elizabeth Laird, Dr. Simon Tew and Professor Peter Clegg

Dr. Olga Vasieva, Prof. Francesco Falciani

Funding

Biotechnology and Biological Sciences Research Council (BBSRC)

Collegiate

Staff and students within the Department of Musculoskeletal Biology, Institute of Ageing and Chronic Disease, University of Liverpool – in particular:

Dr. Mandy Peffers, Ms. Rhiannon Morgan, Dr. Danae Zamboulis, Ms. Yalda Ashraf-Kharaz

Additionally:

Dr. Pernilla Elliasson, Linköping University, Sweden

Mr. Jude Payne, The Scottish Parliament, Edinburgh

Dr. Dagmar Iber, ETH Zurich

Technical support and services

Technical staff within the School of Veterinary Sciences where this project was based.

Contents

Figures

1:

Figures

2:

Figures

Tables

Supplementary Data

3:

Figures

Tables

Supplementary Data

4:

Figures

Tables

Supplementary Data

No data provided

5:

Figures

Standards and Notations

Standard Scientific Units

Standard Scientific Units and symbols are used throughout.

°C	degrees Celsius
cm/mm	centi-, milli-metre
Da	Daltons
g	grams
g	relative centrifugal force
k	kilo

L	litres
µ	micro
m	milli
M	mol(s)
pH	$[H^+]$ at 25°C
ρ	pico
s.d.	standard deviation
V	Volts

Statistical and Mathematical notation

^	exponentiation
Δ	delta, change
ANOVA	analysis of variance
$10^{x/-x}$	Integral exponents. 'E notation' is used for some software outputs

Chemical symbols

Chemical nomenclature, as defined by the International Union of Pure and Applied Chemistry (www.iupac.org), is used where appropriate.

Gene and protein nomenclature

Gene and protein symbols are presented as defined by the Rat Genome Nomenclature Committee. Genes are italicised with the first letter of the symbol in uppercase (*Rps20*) and all subsequent letters in lower case. Proteins follow the same symbol as the gene, but are not italicised and are all in uppercase (RPS20). For human genes and proteins all appear in uppercase with only the gene italicised as defined by the HUGO Gene Nomenclature Committee (www.genenames.org).

Abbreviations

Abbreviations used throughout the thesis may be found here. Where specific foreshortenings are used within a chapter they are presented at the start of that section.

ATP/ADP	Adenosine tri-/di-phosphate
EDTA	Ethylenediaminetetraacetic acid
DEPC	Diethyl pyrocarbonate/diethyl dicarbonate (IUPAC)
FDR	false discovery rate
GO	gene ontology
GMP	Good manufacturing practice
HTML	HyperText markup language
KEGG	Kyoto Encyclopedia of Genes and Genomes
n	independent biological replicates
w/w	mass fraction
w/v	percentage solution
OA	Osteoarthritis
PAGE	Polyacrylamide gel electrophoresis
PBS	phosphate-buffered saline
(q)PCR	(quantitative) polymerase chain reaction
RA	Rheumatoid arthritis
SDS	sodium dodecyl sulphate
U	empirical units

Software Packages

- Where specific functions or algorithms within a software package have been implemented these are recognizable within the context of the methods by the following type-face: `limma`
- Within *R* functions there are many user-defined options, which may be used to optimize the process being undertaken. Only where the default options were not employed will this be stated

Language and Markings

UK English is used throughout, e.g. 'signalling' and not 'signaling'

Latin terms are italicized – *de novo*

References to chapters, figures and sections in the text are in **bold**.

1: General Introduction

1.1: Opening statement

In this introduction an overview is provided of cartilage and tendon anatomy and physiology. The pathophysiological mechanisms of cartilage and tendon degeneration are summarized and the population impact is reported. Current regeneration strategies are presented with explanations of the growth in tissue-engineering research and use of musculoskeletal progenitor cells. The motivations and project objectives are defined relative to this current understanding. Firstly, the philosophy of systems biology and common tools are outlined with view to applying these to quantitative transcriptomic and proteomic data from cartilage and tendon studies.

1.2: Philosophy of Systems Biology

1.2.1: Systems Biology: a paradigm shift in science

To understand the philosophy of systems biology its origins should be considered. Ambiguity and difficulties in finding an adequate definition of systems biology are in part due to the historical development of the discipline. Systems biology is either considered a natural evolution of earlier scientific disciplines e.g. biophysics, and the work of pioneers such as Turing (Turing 1952), or considered a paradigm shift (Marcum 2009, Bard 2013) in scientific thinking. Definitions are also confounded by the complexity of the strategies a systems biology approach encompasses – methods as diverse as, and not restricted to, whole-genome sequencing and 'omics' studies (proteomics, transcriptomics, metabolomics), single-cell microfluidics (Breslauer, Lee et al. 2006), *in silico* modeling, machine learning, and other bioinformatics methods.

Advocates would present systems biology as a Kuhnian paradigm shift, or revolutionary step, in science. According to Kuhn a paradigm shift in science occurs where the recognition of an anomaly in fundamental understanding results in a crises so profound that it propagates a revolution in thinking; this results in a shift in a paradigm toward a competitor (Kuhn 2012). The catalyst for this shift, it has been suggested, was the wave of data generated by the completion of the Human Genome project (Ideker, Galitski et al. 2001, Marcum 2009, Conesa and Mortazavi 2014). Proponents would suggest that traditional reductionist strategies in biology fail to account for the complexity of biological phenomena. Critics, in contrast, would suggest that systems biology represents an evolutionary extension of molecular biology and it certainly builds on the successes of a reductionist methodology (Bard 2013).

Two comprehensive definitions of systems biology, below, give some indication as to the central tenets of a systems approach:

> "Systems biology is a discipline seeking to understand the emergent behavior of a biological system by integrative modeling of the interactions of (all) the molecular elements". (Wang and Levchenko 2009);

and,

> "The goal of systems biology is a predictive understanding of the whole. If the whole is more than the sum of its parts, it follows that acquiring a catalogue of all the parts is not necessarily the first order of business". (Szallasia, Stelling et al. 2010).

1.2.2: The systems biology approach

To define the systems biology approach the ultimate objectives must be considered. Systems biology should be considered as having two goals, firstly to describe how proteins in complex networks work (as proteins drive biological networks) and

secondly, to integrate with the functional or organismal phenotype the molecular and network data (Bard 2013).

In general terms the systems biology approach consists of an iterative series of comprehensive perturbations and systematic quantifications to measure, temporally, elements from all the distinct levels of a biological system. In an attempt to recapitulate the behavior of the system under investigation all the quantitative data must be integrated into a network model. This mathematical model is reconciled against the observed responses then a new hypothesis is formulated and tested experimentally. The model is cyclically refined and perturbed with new findings being incorporated as they arise (Ideker, Galitski et al. 2001, Wang and Levchenko 2009). Systems may be perturbed in a number of ways including high-throughput genetic manipulation, e.g. gene overexpression, systematic gene mutations, RNA interference, small molecule (drug) libraries, and growth hormones. The concomitant response to these perturbations is then quantified as before.

Individual elements of the biological system are not under consideration; rather the relationships between all elements within the system are investigated dynamically. Consequently, systems biology considers all biology as an information science. Discovery investigations (or global profiling studies) are common and these are not necessarily hypothesis-driven at the outset, however, the integration of hypothesis- and discovery-driven approaches is a hallmark of systems biology (Ideker, Galitski et al. 2001). The tools that make systems biology possible include: genomic reference sequences of model organisms, comprehensive gathering of information at all levels of the biological hierarchy, high-throughput, massively parallel and automated methodologies, improved computational power, curated databases, open-source software and multi-disciplinary teams.

What does not constitute a systems biology approach is as relevant as the definitions above. Some researchers do not consider that global data gathering, perturbation of that system and integration of that data alone constitutes systems biology, rather the mathematical modeling of biological systems represents a true systems approach. Sequencing of transcripts from samples in a discovery study does not represent a systems biology approach, however, studies following changes in gene expression in a temporal manner, or in response to siRNAs, may be considered a systems biology methodology (Conesa and Mortazavi 2014).

Complexity in biological systems

Biological phenomena, e.g. development, physiological homeostasis, neoplasia, are complex. The behavior of these phenomena cannot be reduced to a linear representation comprising the sum of the components parts. By extension biological data is complex for several reasons – a) organization operates on multiple, hierarchical levels; b) it is derived from integrated networks; c) these networks are robust to multiple perturbations; d) key target nodes within the network, when perturbed, may have profound effects. Causality is distributed throughout the system and operates in a bidirectional manner between, and within, levels. Furthermore, systems biology assumes that no biological level has preferred status. (Ideker, Galitski et al. 2001, Bard 2013). Complexity is also associated with the fact that data will always be missing, for example failure to identify peptides in mass-spectra or measuring *in vivo* states (rate constants) (Bard 2013), or redundancy in the system, or that biological processes occur over vastly differing time scales. A systems biology approach does not, however, need to include large numbers of elements or data sets and may represent small-scale quantitative modeling, e.g. in yeast studies (Klipp, Liebermeister et al. 2009).

Kitano (2002) defined that a systems-level understanding would require analysis four properties of a biological system, a) the system structure (e.g. the network of gene interactions), b) system dynamics (behavior over time), c) the control method (define the mechanisms that control the state of the cell and modulate them), and d) the design method (principles of design and simulation to modify biological systems to having desired properties) (Kitano 2002)

1.2.3: Systems biology tools

As stated earlier a number of methodologies may be used in a system approach and two of these are described below. Further discussions on quantitative data integration techniques are explored in **Chapters 5** and **7**.

Tools of interpretation in systems biology I: Co-expression network analysis

With the reduction in the cost of microarrays many studies have published lists of differentially expressed genes or gene sets. Prioritised lists, whilst useful reference tools, provide only single gene-specific measurements at the expense of all the data points that did not meet the threshold criteria. Furthermore, it does not consider the interactions of the elements of the system. A systems approach does not make *a priori* assumptions about elements of a system, rather it is explored in a holistic manner. Covariance, a measure of how much two random variables change relative to each other, may be used to establish gene-gene co-expression relationships and may be used to infer patterns in the expression data and create networks of highly interconnected genes (Conesa and Mortazavi 2014). Established systems approaches, such as weighted gene co-expression network analysis (WGCNA), can be applied to many biological contexts to define modules of highly co-expressed genes and derive phenotypic associations from them (Langfelder and Horvath 2008, Gaiteri, Ding et al. 2014). This will be explored further in **Chapters 4** and **5**.

Pathways are graphical representations of models encoding the interactions between genes, proteins and metabolites within cells, tissues and organisms – pathway topology is the overall arrangement of the elements of this model. The production of lists of differentially expressed genes between two conditions has now become relatively routine, however, predicting the phenomena that drive a phenotype is still challenging. It has been suggested that the ability to infer, correctly and consistently, the pathway perturbations promoting a phenotype from a list of genes may be what translates vast data gathering exercises into meaningful biological knowledge (Mitrea, Taghavi et al. 2013).

In order to make inferences about networks the properties of the system, mechanistic and structural, must be mapped. These networks have a complex topology, which, although likely to be biologically relevant, may be difficult to interpret in terms of cellular function (Conesa and Mortazavi 2014). Furthermore, causality in biological systems is bidirectional within, and between, hierarchical levels (Marcum 2009, Bard 2013), therefore profiling only gene expression is not consistent with a systems approach.

Long lists of differential abundance provides no mechanistic insight into the system under investigation (Khatri, Sirota et al. 2012). Pathway analysis is a group of methodologies that seeks to gather cohorts, or sets, of genes that function within the same pathways; achieving this reduces the complexity of the data from thousands of genes to scores of pathways. Rather than producing a list of pathways this type of analysis should define the activity status of pathways thereby improving our ability to explain the data.

The term 'pathway' is often misappropriated, for example, gene ontology terms do not describe a pathway, nor do protein-protein interaction networks (Khatri, Sirota et al. 2012). Most pathway analysis is driven through the use of a pathway knowledge base, which describes a process, structure, or components. Gene expression patterns for the condition under review are correlated with information within a pathway knowledge base. Numerous techniques are available, though most are derivative and use overlapping resources. The most widely used and accessible method is 'over-representation analysis' (ORA) where the statistical evaluation of a subset of genes within a differential expression list matching to a particular pathway is considered. Over-representation analysis assumes each gene, or pathway, is an independent entity, contrary to the understanding of the complex interactions that exist between them, and so limits any insight into the pathways.

Functional class scoring (FCS) techniques, such as Gene Set Enrichment Analysis (Broad Institute, http://www.broadinstitute.org/gsea), try to address some of these issues by not having arbitrary thresholds, using molecular measurements, and consider coordinated changes between genes. However, pathways are still considered independently and many methods may rank and discard genes in a particular pathway. For both methods, described above, as only the number of genes or co-expression of those genes is used to define significant pathways were these pathways to be redrawn with entirely different links between the genes both ORA and FCS would produce the same results. As lists of genes, or 'gene-sets' do not include any additional information on a pathway they cannot be considered pathway analysis (Tarca, Draghici et al. 2009, Mitrea, Taghavi et al. 2013).

Pathway topology methods use many of the same steps as functional-class scoring, however, the gene-level statistic is calculated on pathway topology. Pathway topology

makes use of the previously collected knowledge about the structure of a pathway, in addition to the molecular measurements. An impact factor analytic approach is a more recent development; this takes into account not only the changes in gene expression, but also the positions of genes in pathways and types of interactions. The impact factor of a pathway is the sum of perturbation factors for all the genes in a pathway. In **Chapter 2** this methodology, implemented through the SPIA algorithm (Tarca, Draghici et al. 2009) was employed to predict the activation status of KEGG canonical pathways based upon the Illumina gene expression data sets. This method is becoming more widely used (Balbin, Prensner et al. 2013, Nance, Smith et al. 2014, Nouailles, Dorhoi et al. 2014), however, is already likely to have been superseded by a number of new methods using several curated databases, implementing multivariate analysis and incorporating multi-dimensional genomics data (Vaske, Benz et al. 2010, Mitrea, Taghavi et al. 2013). However, many are limited by their ease of implementation.

Working definition of systems biology

In the context of this thesis a systems biology approach is understood to be the collection of methodological steps, which together aim to generate novel and testable hypotheses about the system under investigation. Initially, hypothesis-free discovery projects survey and quantify global transcriptome and proteome profiles for a series of conditions using cartilage and tendon. This data is curated and integrated with functional *in vivo* and *in vitro* phenotypes at the network level to define emergent properties of the system. Protein-protein interaction networks arising from these relationships are used to generate a hypothesis about regulatory units driving the phenotype of interest with the intention of perturbing these units in future and feeding into computational models.

1.3: Structure & Function of Cartilage and Tendon

1.3.1: The musculoskeletal system: an introduction

The musculoskeletal system consists of a diverse group of tissues that are developmentally, anatomically, and functionally inter-related. Bone confers the structural rigidity required for locomotion; muscle contraction generates force, tendon transfers that force, whilst cartilage provides a smooth surface for unimpeded angular movements at joints. Contributing to this mechanical system are ligaments, synovial tissue that lines joints and provides lubricating fluid, and complex extracellular matrices. Of these tissues cartilage and tendon are the focus of the studies presented in this thesis.

To demonstrate the common ontogeny, disease pathophysiology, and regeneration strategies for cartilage and tendon three sections provide an overview: i) structure and function; ii) disease and development; and iii) repair and regeneration. In each of these the key issues highlighted and how they inform and influence the motivations for the series of studies presented in this thesis. A final section clarifies the core objectives of the thesis and the novelty of the study design.

It is not possible within the constraints of this thesis to fully survey the literature pertaining to development, anatomy, pathophysiology, regeneration and therapeutics for each tissue and these disciplines will be considered in general terms. Where points of interest are expanded upon elsewhere in the thesis these will be referenced in the text.

1.4: The extra-cellular matrix

As the majority of both cartilage and tendon consists of specialised extra-cellular matrix (ECM), and contribute to a loss of function in disease, an overview is provided.

Although the ECM is not the principal focus of this thesis it is fundamental to an understanding of the goals of musculoskeletal regeneration strategies. The ECM in crude terms is often considered the filler between cells in all tissues, but in reality represents both a physical scaffold and a transducer of biochemical and biomechanical cues with roles in morphogenesis, differentiation, homeostasis and is implicated in variety of disorders related to age, injury, or neoplasia (Frantz, Stewart et al. 2010). Although the fundamental components are comparable across many tissues the proportional composition is functionally related and unique to each tissue; it also exists as the interface between tissues and between cell populations within a tissue.

The ECM is a dynamic structure involved with sequestration and release of growth factors. It undergoes constant remodeling, and interacts with the resident cells through integrins, syndecans, and discoidin domain receptors. Excluding water, the ECM components are proteins, glycoproteins or proteoglycans. The former are, in various proportions: fibrous collagens, elastins, laminins, fibrillins, thrombospondins, fibulins, tenascins and fibronectin (Halper and Kjaer 2014). The latter, proteoglycans, consist of glycosaminoglycans (GAGs) with a covalent link to a core protein; this core protein is not present in hyaluronan. The composition of these structures may classify proteoglycans into: cell surface-associated (syndecans, glypicans), small leucine-rich proteoglycans or SLRPs (decorin, lumican), or modular (aggrecan, versican, perlecan). In general, proteoglycans are highly hydrophilic, linear molecules that confer the resistance to compressive forces through the formation of hydrogels (Schaefer and Schaefer 2010). Furthermore, each proteoglycan group has relevance to signalling modulation, cell adhesion, migration and proliferation (Frantz, Stewart et al. 2010, Schaefer and Schaefer 2010).

The CCN-family of proteins is of particularly relevance to ECM function. The six secreted CCN proteins (CYR61, CNN2/CTGF, CNN3/NOV and WISP 1-3) have specific ECM associations with roles as adaptor molecules linking the intracellular state and the ECM signalling through integrins and proteoglycans. They are induced by growth factors and cytokines including TGF-β and are also implicated in cell differentiation, chondrogenesis, proliferation and migration (Leask and Abraham 2006).

1.4.1: Collagens

Collagens represent one third of the human proteome (Frantz, Stewart et al. 2010, Ricard-Blum 2011, Chang, Shefelbine et al. 2012) and collagen alpha-helical chains, which represent the fundamental unit of the ECM, intertwine with other collagen helices to form homo- or hetero-trimeric collagen helices contributing to the structural integrity of tissues. Each collagen chain contains repeating glycine (Gly,G)-X-Y triplets, of which X and Y are usually proline (Pro,P) or hydroxyproline (Hyp,O) (van der Rest and Garrone 1991, Ricard-Blum 2011). Triple-helical collagenous domains are flanked by non-collagenous regions; these often containing recognisable domains found in other matricellular proteins. Collagens may be loosely grouped into: fibrillar, fibril-associated, network-forming, transmembranous, endostatin-precursors, or within a miscellaneous group (Gordon and Hahn 2010). For the purposes of this introduction only the fibrillar collagens will be considered.

The fibrillar collagens are types I (the must abundant in tendon), II (the most abundant in cartilage), III, V and XI. Collagen type I principally exists as a hetero-trimer consisting of two alpha-1 and one alpha-2 collagen chains; homo-trimeric forms of collagen type I, three alpha-1 chains, exist in foetal tissue, musculoskeletal disorders (Dupuytren's contracture and osteogenesis imperfecta) and neoplastic stroma with

altered structural and mechanical properties (Chang, Shefelbine et al. 2012). In contrast, collagen type II exists only as a homo-trimeric molecule (Gordon and Hahn 2010).

Pro-collagen biosynthesis

Pro-collagen units are the precursors of collagen and undergo conversion, complex post-translational modifications and trafficking through an intra-cellular secretory pathway to ultimately result in the deposition of collagen fibrils into the ECM. In the pro-form triple helices are flanked by N- and C-terminus pro-peptide globular extensions, which undergo proteolytic cleavage by N- and C-proteinases respectively (Kadler, Baldock et al. 2007). The removal of the pro-peptides permits the aggregation of collagen triple helical molecules into fibrils.

Pro-collagen chain biosynthesis occurs in the endoplasmic reticulum (ER) and nucleation of three pro-collagen monomers into the triple helix occurs at the C-pro-peptide under the influence of various enzymes with resultant post-translational modifications (PTM) (Banos, Thomas et al. 2008). For example, the conversion of proline residues (in the Gly-X-Y repeating triplets of the collagen chains) to hydroxyproline is performed by prolyly-4-hydroxylase, with ascorbic acid (Vitamin C) as an essential co-factor (Canty and Kadler 2005). Trafficking covers the transition of collagen triple helices from the ER to the plasma membrane, via the Golgi, for secretion from the cell. Vesicular transport clusters transport pro-collagen to the *cis*-Golgi after budding directly from the ER membrane (Banos, Thomas et al. 2008). In this model the cleavage of the C- and N-globular domains occurs intracellularly; this is the final PTM and creates the highly insoluble tropocollagen rods that may spontaneously self-assemble to form fibrils (Banos, Thomas et al. 2008). Proteinase activity is specifically mediated by C- and N-propeptide tolloid and ADAMTS family enzymes respectively (Canty and Kadler 2005), although these have other substrates additionally.

1.4.2: Integrins coordinate chondrocyte communication with the ECM

The chondrocyte is isolated within the general ECM and peri-cellular matrix (see **Chapter 3**), lacking cell-to-cell contacts and is supremely sensitive to changes in the microenvironment – it is connected to it by integrins and these influence chondrocyte responses (Demoor, Ollitrault et al. 2014), i.e. 'integrating' the extracellular matrix, cytoskeletal components and signalling pathways. Integrins are heterodimeric transmembrane proteins consisting of α and β subunits that have extracellular domains defining the matrix ligands (Loeser 2014). Of the 24 integrin heterodimers described chondrocytes express seven (α1β1, α3β1, α5β1, α10β1, αVβ1, αVβ3, and αVβ5) in physiological circumstances, with elevated levels of α1β1 and α3β1 and novel heterodimer expression (α2β1, α4β1) evident in osteoarthritis (Almonte-Becerril, Costell et al. 2014, Loeser 2014). Integrins can recognize distinct collagen subgroups, for example α10β1 is collagen, type II binding and is limited to cartilage (Heino 2014). Integrins function to mediate cell-adhesion to ECM proteins in a substrate-restricted manner, e.g. fibronectin, collagen type II. Binding stimulates 'outside-in' signalling networks that converge on mitogen-activated protein (MAP) kinase family of proteins (ERK, JNK, p38) to influence down-stream transcription of genes through NF-κB and AP-1 (Fos-Jun dimer). Canonical integrin signalling results in the rapid increase in levels of phosphatidylinositol lipid messengers (PI-3K/Akt signalling, see **Chapter 3**) and receptor tyrosine kinase-mediated phosphorylation of proteins including focal adhesion kinase (FAK) and Rho/Rac activation (Heino 2014, Loeser 2014). In general, the activity of integrins has been shown to include: mechano-transduction, survival, differentiation, and proliferation. Additionally, matrix fragments, especially fibronectin, influence the catabolic signalling through RGD motif binding by integrins including α5β1, resulting in a signalling cascade promoting increased expression of MMPs, pro-inflammatory cytokines (PGE2/COX2), reactive oxygen species (ROS) and other

degradative enzymes (ADAMTS5). The generation of further matrix protein fragments can perpetuate a cycle of degenerative effects on cartilage (Loeser 2014).

Summary

The ECM is a dynamic and complex structure that confers the functional mechanical properties of cartilage and tendon. Dysregulation of its production and degradation are critical to the pathophysiology of musculoskeletal disorders.

1.5 Cartilage

> *"Where-ever the Motion of one Bone upon another is requisite, there we find an excellent Apparatus for rendering that Motion safe and free."* (Hunter 1743)

1.5.1: General anatomy

Having defined the principal structural components of cartilage and tendon each tissue will be considered in turn with regard to the general anatomy and function.

Adaptations of cartilage may be found wherever, in the axial or appendicular skeleton, two surfaces are required to move over one another without impedance. Cartilage may be described as: fibro-cartilage, elastic, fibro-elastic or hyaline (Eyre 2002). Articular cartilage, a smooth, load-bearing surface that allows frictionless motion of a joint, is a form of hyaline cartilage. In the healthy adult articular cartilage is devoid of blood and lymph vessels, and is aneural. The only cell-type is the chondrocyte, which is responsible for the synthesis of extracellular matrix and maintenance of the cartilage architecture. These critical cells only account for 1-5% of the cartilage structure. Nutrition is conferred by diffusion and chondrocytes have no cell-to-cell contacts (see integrins, below). The physiological environment is considered to be hypoxic as a result

of the low oxygen tension at the level of the chondrocyte; consequently anaerobic glycolysis is the main mechanism for ATP generation (Demoor, Ollitrault et al. 2014). As wet weight, over two-thirds of cartilage is water; this accounts for the load-dependent deformation. Collagens represent a further 10-20% of cartilage volume, of which collagen type II is the most prevalent. The principal collagens in mammalian cartilage (types II, IX and XI) exist as cross-linked collagen hetero-polymers to form, at the ultra-structural level, a random network of fibrils, in contrast to tendon (see below). Other collagens, type III, VI and X are also present, with type X restricted to the calcified cartilage interface with the sub-chondral bone (Eyre 2002, Bhosale and Richardson 2008, Demoor, Ollitrault et al. 2014). Four zones are apparent in articular cartilage by light microscopy. Passing deep to the articular surface these are: superficial/tangential, intermediate/transitional, deep/radial and calcified. These layers differ in their collagen fibril orientation, which, along with the extensive collagen cross-linking ensures the material strength of cartilage (Eyre 2002).

Non-collagenous matrix and proteoglycans

Protein polysaccharides, proteoglycans (mainly aggrecan), account for an additional 10-20% of cartilage volume and contribute to the compressive strength of cartilage. In addition to aggrecan other aggregating proteoglycans, such as versican, form large multi-molecular complexes with hyaluronan, an interaction that is stabilised by hyaluronan link protein (HAPLN1). Small leucine-rich proteins (biglycan, decorin, lumican, fibromodulin), which interact with fibrillar collagens, and miscellaneous proteoglycans (perlecan, lubricin/PRG4) are also components of the cartilage ECM (Roughley 2001). Of the non-collagenous proteins a number of structural components are present including: cartilage oligomeric matrix protein (COMP), matrilin-1/cartilage matrix protein (MATN1), thrombospondins (THBS1, 3, 4), fibronectin (FN), chondroadherin

(CHAD), elastin (ELN) and fibrillin (FBLN). Disruption of these matrix components, either through congenital mutations, for example COMP in skeletal dysplasias (Hecht, Hayes et al. 2005), or acquired mechanisms, e.g. proteolytic activity (Mort and Billington 2001), can contribute to the degradation of cartilage and loss of function.

1.6: Tendon

1.6.1: General anatomy

Tendons are the functional link between the static and dynamic parts of the musculoskeletal system transferring the forces generated by muscular contraction to the skeleton and, thus, facilitating motion. Following injury this function is impaired. Parity of functionality with the pre-injury state after healing is rarely achieved; the final tensile strength of healed tendon has been reported as being reduced by up to thirty percent (Müller, Todorov et al. 2013).

Comprising dense bundles of parallel collagen fibres in close apposition confers tendons with specialist mechano-transducive properties of high tensile strength. Elasticity is a function of the elastin molecule tropoelastin, which aggregates and is stabilized by cross-links in a lysyl-oxidase-dependent manner (Scott 2003). As a caveat to the statements above tendons may also exist as 'intramuscular tendons', facilitating pennation of muscle bellies, or as 'intermediate tendons' connecting two muscle bellies (Benjamin, Kaiser et al. 2008). Furthermore, although in general tendons act to transmit tensile forces derived from skeletal muscle in some anatomical locations forces of compression and shear are also applied, for example, where tendons pass over a curved region of bone and act as a pulley. Other tendons, such as the highly developed equine superficial digital flexor tendon, have the capacity to store elastic energy for efficient

locomotion (Thorpe, Udeze et al. 2012). Some anatomical features are defined by the presence of calcified structures or areas within the tendon including distal limb and pedal sesamoid bones in the horse and man, and areas of fibro-cartilaginous adaptation at tendon insertions (Benjamin, Kaiser et al. 2008).

The canonical tendon structure, as described by Kastelic and co-workers (1978) (Kastelic, Galeski et al. 1978) consists of hierarchically organized structural units of insoluble collagen type I molecules aggregated to form collagen fibrils; subsequently these are collected to form 'fibres' – the basic functional unit of the tendon. Primary, secondary and tertiary fibre bundles form from sequential aggregation. Combined tertiary bundles form the tendon, which is surrounded by the epitenon connective tissue, blood vessels and nerves (Kannus 2000).

Tenocytes, or tendon fibroblasts, are the majority cell type in an otherwise sparsely cellular tissue, and are responsible for the production and secretion of ECM and collagen turnover. As dry weight, tendon is predominantly comprised of collagen of which type I collagen represents 95% with the remainder consisting of type III, V, XII and XIV (Müller, Todorov et al. 2013, Thorpe, Birch et al. 2013). Within each level of the tendon hierarchy small amounts of non-collagenous matrix are present (Kastelic, Galeski et al. 1978), which has been relatively ill-defined. The tendon non-collagenous matrix (NCM) comprises glycoproteins including COMP, lubricin and tenascin-C and other proteoglycans, mainly SLRPs, with decorin the most abundant. The relative abundance of the NCM is associated with the functional requirements of the tendon with areas of high compressive load requiring higher concentrations of NCM, whilst areas with high tensile loads are sparse by comparison (Thorpe, Birch et al. 2013).

Summary

These cursory reviews of tendon and cartilage anatomy and structure demonstrate a commonality in terms of function defining structure, sparse cellularity, and a preponderance of collagenous matrix and varying proportions of non-collagenous matrix. This commonality extends to the morphogenesis of each tissue as outlined in in the following section.

1.7: Disease & Development

1.7.1: Common origins: models of musculoskeletal development

Cartilage and tendon progenitors

An understanding of cartilage and tendon morphogenesis, through the limb bud developmental paradigm, provides a conceptual framework within which an understanding of the regulatory elements and networks driving pathophysiological and regenerative mechanisms may be developed.

Whereas muscle progenitors arise from the myotome, cartilage, bone, tendon and ligament all arise from undifferentiated cells within the sclerotome, lateral plate mesoderm or neural crest (Sugimoto, Takimoto et al. 2013). For the limbs, progenitors are derived from the lateral plate mesoderm. At appropriate topographical regions, encoded by homeobox (*Hox*) genes (Zakany and Duboule 2007), the progenitors displace from the lateral plate mesoderm and assemble beneath ectodermal pockets forming dome-like structures called 'limb buds'. As these buds rapidly grow invasion by the primitive muscle and vasculature cells occurs. In the classical model proximo-distal outgrowth of the limb bud is promoted by the apical ectodermal ridge (AER), which is relevant to our understanding of musculoskeletal morphogenesis because it produces a

number of signals relevant to specification of mesenchymal progenitors, particularly fibroblast growth factor (FGF) family members (Zeller, López-Ríos et al. 2009). The ultimate fate of skeletal progenitors is defined by sequential, modular signals generally consisting of the bone morphogenetic proteins (BMP), transforming growth factor beta (TGF-β), fibroblast growth factor (FGF) and Wingless/Wnt and Hedgehog family of growth factors (Lorda-Diez, Montero et al. 2014) through the generation of dynamic spatial and temporal morphogen gradients (Dekanty and Milan 2011). It is these local signals that drive the divergent differentiation of mesenchyme to cartilage and tendon. It is proposed that as the limb bud proliferates the distal most mesenchymal cells remain in an undifferentiated state whilst core mesenchyme, no longer under the influence of defined morphogen gradients, activate SRY-box containing-9 (*Sox9*) expression to initiate chondrogenic differentiation.

Summary

Common mesenchymal progenitor cells differentiate into many musculoskeletal tissues with diverse functions. Growth factor gradients and homeodomain genes coordinate and spatially restrict development of nascent cartilage and tendon structures. An understanding of developmental ontogeny is critical to understanding directed differentiation in regenerative therapies and biomimetic cultures.

1.7.2: Chondrogenesis, articular cartilage and joint development

The condensation of mesenchymal progenitors to form discrete, self-organising cartilage templates (anlagen) is the initial process in skeletogenesis. Associated with this transition is a move from expression of collagen types I, III, and V to cartilage-associated collagens II, IX, and XI. Proliferating chondrocytes express collagen VI and matrilin-1 (MATN1) under the control of the PTHrP/Ihh axis (Goldring 2012). Cartilage determinism, and the prevention of differentiation towards an osteogenic

lineage, is under the control of a number of signalling pathways and transcriptional regulators. Activation of chondrogenic pattern formation may be under the influence of TGF-β signalling, possibly via a reaction-diffusion mechanism (Turing 1952, Miura and Shiota 2000, Christley, Alber et al. 2007). Complex interplay between TGF-β, BMP- (Pizette and Niswander 2000, Yoon, Ovchinnikov et al. 2005) and Wnt-(Day, Guo et al. 2005, Hill, Später et al. 2005) signalling all contribute to the balance between osteo- and chondro-genesis. The major transcriptional regulators are RUNX2 and SOX9 for osteogenesis and chondrogenesis respectively (Goldring 2012). SOX9, with the addition of SOX5 and SOX6, are a triad of transcriptional controllers that cooperatively activate collagen type II expression (Lefebvre, Li et al. 1998, Lefebvre, Behringer et al. 2001).

More is understood about the regulatory elements involved in chondrogenic determinism in skeletogenesis than definition of articular cartilage formation. The complex three-dimensional architecture of joints and the compound nature of the tissues that comprise them (cartilage, ligament, synovium, meniscus, tendon) are not well described in terms of the pool of progenitors from which they arise. Decker, *et al* (2014) have proposed a model for the emergence of joints, and formation of articular cartilage, from undifferentiated mesenchyme based upon a review of a number of genetic cell lineage tracing studies (Decker, Koyama et al. 2014). The authors proposed that SOX9/COL2A1/doublecortin-positive cells within the cartilaginous anlagen, which is linear and uninterrupted in early embryogenesis, becomes segmented by unknown mechanisms; $GDF5^{+}$ cells, which define the avascular mesenchymal 'interzone' that interrupts the adjacent cartilaginous elements, together with migrating cells give rise to the articular cartilage, synovial lining and intra-joint ligaments. Matrillin 1-expressing chondrocytes residing within the anlagen diverge to undergo endochondral ossification,

whilst matrillin-1 negative cells with COL2A1-expression history go on to differentiate into articular chondrocytes. Furthermore, TGF-βR2-expressing cells would emerge as slow-cycling, reserve progenitor cell population.

1.7.3: Tendon development

The development of the embryonic tendon demonstrates heterotypic induction through the intimate association with the developing structures of the limb, in particular the developing muscle (Aslan, Kimelman-Bleich et al. 2008), but also developing cartilage (Schweitzer, Zelzer et al. 2010). These complex interactions are required to ensure the robust anchoring of tendon to muscle at the myotendinous junction and to bone at insertional entheses without which force transduction would not be possible. There is evidence that axial tendons arise from a different embryonic lineage than the developmental mechanisms that promote appendicular, or limb, tenogenesis thus further complicating our understanding of potential regenerative processes (Brown, Finley et al. 2014).

Regulators of tendon differentiation

The development of tendon repair mechanisms is partly limited by our understanding of tendon development, the paucity of data on tendon differentiation cues, and evidence of the role of biomechanics in development. Some lineage-restricted transcription factors have been implicated in tendon formation in a time-restricted manner with scleraxis and SOX9 associated with the emergence of progenitors, whilst Mohawk (MKX) (Anderson, Arredondo et al. 2006, Anderson, Beres et al. 2009) and early growth response proteins EGR1 and *2* (Lejard, Blais et al. 2011, Guerquin, Charvet et al. 2013) are associated with differentiation and maturation stages (Liu, Zhu et al. 2014).

The most notable of the tendon-specific mediators is scleraxis (SCX), a basic helix-loop-helix (bHLH) transcription factor, which was detected in mouse embryos from 9.5 days *post coitus (*d.p.c.*)* with high levels demonstrable in cartilage, tendon and ligaments (Cserjesi, Brown et al. 1995, Schweitzer, Zelzer et al. 2010). For axial tendon development tendon-progenitor cells are defined by the expression of SCX, in early developmental stages, within the syndetome compartment arising from the interaction of the somitic compartments of the sclerotome and the myotome (Brent, Schweitzer et al. 2003). The former giving rise to the dorsal dermis and axial skeleton, whilst the latter to skeletal muscle. Fibroblast growth factor signalling from the adjacent myotome induces and propagates development of the SCX-positive cells. In comparison, the SCX^+ cells in the developing limbs arise from superficial dorsal and ventral aspects of the limb bud mesenchyme without the compartmentalization described for axial tendon (Brown, Finley et al. 2014). Here interactions between bone morphogenetic proteins, and the BMP antagonist Noggin, were reported to influence the domain of SCX^+ cells (Schweitzer, Chyung et al. 2001).

In general the molecular targets considered for neotendon formation have arisen from our understanding of tendon development, but largely consider the tendon in isolation belaying what is clear from tendon ontogeny, that tendon development relies on the cross-talk of muscle, bone and cartilage (Schweitzer, Zelzer et al. 2010). Additionally, evidence for the tenogenic effects of these targets is limited. Putative targets for tendon regeneration include SIX1 and SIX2, SMAD8, TGF-β and BMP-family members including GDF5, GDF6 and GDF7 (Aslan, Kimelman-Bleich et al. 2008).

In additional to soluble factors, such as the role of TGF-β signalling in the recruitment and maintenance of tendon progenitors (Hasson 2011), mechanical factors are known to have roles in the development and homeostasis of tendon function. Tissue

engineering studies have considered how best to utilize mechanical cues, such as substrate stiffness (elastic modulus) or to provide guides for scaffolds and bioreactors. Studies considering the mechano-activity of developing embryonic cells have been few and conflicting in their conclusions. Using various forms of neuromuscular blockade several studies have demonstrated the impact of altered skeletal muscle contraction on the developing tendon with reduced, degenerative, or inhibited tendon structure formation, and reductions in the expression of tenascin-C or numbers of SCX^+ cells (Schiele, Marturano et al. 2013). In the absence of muscle, early tendon progenitor distribution and tenogenic induction is not inhibited in the developing avian limb, however further progress is retarded at the development and maturation stages and primitive tendons degenerate (Kardon 1998). For the developing axial tendon there is no muscle-independent stage documented (Brown, Finley et al. 2014). Whether this muscle dependence is derived from mechanical cues, soluble factors from muscle, or a combination of both, still requires further study (Schiele, Marturano et al. 2013).

Summary

The induction of tendon development does not occur in isolation and requires coordinated interactions including muscle and cartilage. Key regulators of tendon development included SCX, MKX and SOX9. Any consideration of the development of tendon regeneration strategies should also consider an integrated understanding of musculoskeletal development.

1.7.4: Defining cartilage and tendon progenitors

Cartilage and tendon progenitors in the limb have been defined principally by the differential expression of the key regulators SOX9 and SCX respectively. This assertion that scleraxis-positive cells alone form tendons in the limb bud has been based upon the continued expression of SCX rather than the absolute tracing of SCX-positive cells

from common progenitors to functional tissue. In early mouse studies the close temporal and spatial relationship of SCX and SOX9 was found in mesenchymal progenitor cells; these cells diverged at E11.5 d.p.c. into cells with restricted expression to form tendon and cartilage respectively (Asou, Nifuji et al. 2002).

More recent studies, however, suggest that cell-fate determination on the basis of categorical expression of either SCX or SOX9 is insufficient. Takimoto, *et al* (2012) demonstrated the direct conversion of tenomodulin (TNMD)-expressing tenocytes into chondrocytes by the forced expression of SOX9 alone (Takimoto, Oro et al. 2012). Further work demonstrated that tenocytes are derived from both $SCX^+/SOX9^+$ and $SCX^+/SOX9^-$ cells; the proportion of the former expression profile increased the closer the tenocytes came to the nascent cartilage (Sugimoto, Takimoto et al. 2013). Conditional inactivation of SOX9 in $SCX^+/SOX9^+$ resulted in the defective formation of the junction between tendon and cartilage. The role of a specialist population of $SCX^+/SOX9^+$ in forming the enthesis, the junction between bone and tendon, has also been demonstrated (Blitz, Sharir et al. 2013). In lineage tracing studies $SOX9^+$ cells were indicative of precursors for articular and growth plate cartilage, in addition to ligament and tendon (Soeda, Deng et al. 2010). The results from this study, suggested the authors, were consistent with part of the nascent tendon cells arising from a contribution by $SOX9^+$ cells of the condensed mesenchymal cells of the cartilage primordia.

Summary

Recent evidence points toward specialist sub-populations of tendon and cartilage progenitors that enable formation of transitional zones between developing tissue, furthermore, it is clear that chondrogenic cells make some contribution to the developing tendon. An understanding of the common origin and intimate spatial

development of cartilage, tendon, and muscle is valuable to the development of organotypic models and bioengineered tissue.

1.7.5: Musculoskeletal disease: population impact

Disorders of the musculoskeletal system are highly prevalent comorbidities of aging populations in contemporary society (Siebens 2007, van Dijk, Veenhof et al. 2008). Debilitating in their own right musculoskeletal disorders contribute significantly to the global burden of disease, the fourth most prevalent (Murray, Vos et al. 2012), have a wider impact on rehabilitation of parallel pathologies (obesity, strokes, cardiovascular disease) and so represent a major health policy issue. Despite this population impact musculoskeletal research and clinical trials are not necessarily proportionately funded or investigated (Bourne, Whittle et al. 2014, Rankin, Sprowson et al. 2014). Medical interventions aimed at management of musculoskeletal tissues vary from physiotherapy, arthroscopy, anti-inflammatories, immunomodulation and numerous 'regenerative' therapies, how there is no treatment for conditions such as osteoarthritis. The basis for complex musculoskeletal diseases, such as osteoarthritis, is multifactorial. Although heritable factors account for 50% of an individual's risk of developing osteoarthritis only 16 disease risk loci have been consistently identified (Panoutsopoulou and Zeggini 2013).

1.7.6: Cartilage Disease Pathogenesis

> "*If we consult the standard Chirurgical Writers from Hippocrates down to the present Age, we shall find that ulcerated Cartilage is universally allowed to be a very troublesome Disease…and that, when destroyed, it is never recovered.*" (Hunter 1743)

This statement, above, by the 18th century surgeon William Hunter was prescient in its forecasting. Over 250 years later the therapeutic strategies available to treat osteoarthritis (OA) range from benign neglect and anti-inflammatories to total joint replacement. As discussed earlier the inherent capacity of cartilage to respond to injury is limited. Globally, osteoarthritis of the hip and knee joints is one of the leading causes of disability and a highly prevalent comorbidity of the ageing society alongside Type II diabetes and neurological conditions. Of the 291 conditions surveyed in the Global Burden of Disease Study (2010) hip and knee OA ranked as the 11th highest contributor to disability (Cross, Smith et al. 2014), accounting for 6.8% of the total disease burden (Murray, Vos et al. 2012). Adjusted for age, the global prevalence of osteoarthritis was 3.8% and 0.85% for the knee and hip respectively (Cross, Smith et al. 2014). This places musculoskeletal disorders high on the ranking of major health policy issues, however, funding levels and research activity globally lag behind other fields (Bourne, Whittle et al. 2014, Rankin, Sprowson et al. 2014)

Osteoarthritis cannot be considered a single disease with a linear narrative to describe the pathogenesis, rather it is a heterogenous condition with multiple causation, where joint failure represents the common end-point (Cicuttini and Wluka 2014). Where the adage 'abnormal wear on normal cartilage; normal wear on abnormal cartilage' is a useful lay description of the pathogenesis in reality there are a myriad of risk factors contributing to chondral damage. It is beyond the scope of this review to cover the plethora of disorders arising from, and associated with, degenerative and inflammatory conditions of cartilage. In **Chapter 5** a review of recent studies of the genetic associations with osteoarthritis will be presented.

Cartilage degeneration: balance of anabolic and catabolic process

The chondrocyte is solely responsible for the turnover of cartilage ECM, however, articular chondrocytes are post-mitotic in the adult and turn-over is very slow – chondrocytes therefore regulate the synthesis and degradation of ECM molecules through the fine balance of catabolic (matrix metalloproteinases, aggrecanases, IL-6) and anabolic (IGF-1, BMP, TGF-β, TIMPs, PDGF) components (Demoor, Ollitrault et al. 2014). Degeneration of cartilage in OA is defined by the degradation of the ECM by a variety of catabolic enzymes including matrix metalloproteinases (MMPs) and aggrecanases in response to elevated levels of interleukin-1β (IL-1β) or tumour necrosis factor-α (TNF-α). These pro-inflammatory factors establish an intra-articular and intra-cellular milieu that disrupts the narrow homeostatic range chondrocytes try to preserve (Demoor, Ollitrault et al. 2014). This establishes chondrocyte dedifferentiation, the loss of the functional phenotypic hallmarks, through the elevated expression of type I collagen (Young, Smith et al. 2005).

1.7.7: Tendon disorders

Within the Global Burden of Disease 2010 study disorders relating to tendon were considered within 'other musculoskeletal disorders' as sequelae to injury of tendon and ligaments are not often part of long-term data capture (Smith, Hoy et al. 2014). As this analysis also included a variety of rheumatoid and musculoskeletal conditions an accurate estimate of the burden of disease attributable to tendon and ligament injury is not possible. If the incidence of acute trauma alone is considered 130,000 patient per annum are presented in the USA with injuries related to the Achilles tendon, rotator cuff of the shoulder, or patellar tendon (Sakabe and Sakai 2011); elsewhere a total incidence for tendon-associated injuries of 52.1 – 166.6/100,000 per year, for men and women respectively, has been reported (Clayton and Court-Brown 2008). It is evident

that a lack of contemporary epidemiological data is available to represent the true incidence of tendon and ligament associated disorders.

Tendon healing

The sub-optimal mechanical integrity of the repaired tendon is a key issue in the return to athletic function. In the acute injury scenario healing of tendon may be organized into distinct phases, which in total may require several months before normal physiological loading may be introduced. Immediately following trauma, bleeding and clotting in the tendon defines the 'haemorrhagic state'. Initial platelet degranulation, releasing cytokines and growth factors, is succeeded by the infiltration of the injury site with neutrophils and macrophages. In the following week a 'proliferative stage' consisting of the migration and proliferation of extrinsic (peritendinous soft tissue, fascia, periosteum) and intrinsic (epitenon, endotenon) cells at the site of injury form a delicate granulation tissue comprising mainly type III collagen. Approximately one month after acute injury the 'formative phase' consists of the active resorption and production of collagen by intrinsic fibroblasts mainly from the epitenon. As this nascent tissue matures there is a gradual reorientation of fibres longitudinally in accordance with the prevailing tensional forces. The 'remodelling phase' is defined by the return to physiological loading imparted by the maturing tissue reaching maximal biomechanical strength. The collagen fibres become more organized in the longitudinal plane, and there is more cross-linking. The type III collagen from the formative phase is slowly replaced by the more mechanically resistant type I collagen, however, this healed tendon callus remains hypercellular, with a greater proportion of type III collagen fibres relative to the uninjured state and with thinner collagen fibres (Voleti, Buckley et al. 2012, Connizzo, Yannascoli et al. 2013, Müller, Todorov et al. 2013). This

failure to recapitulate the native state may account for the high morbidity seen with tendon injury.

Conflict in pathophysiology of tendon disorders

Tendon disorders encompasses a range of clinical presentations from acute rupture to chronic degenerative tendinopathy. Strain-induced injury in energy-storing tendons (e.g. Achilles and patellar tendons) is the most common form of tendon injury in athletic humans. Unlike positional tendons, energy-storing tendons serve to transmit forces generated by muscles and store elastic energy as a mechanism for the conservation of energy (Thorpe, Udeze et al. 2012). Both age and level of athleticism have a strong association with the incidence of strain-induced tendinopathies (Smith, Birch et al. 2002). The underlying pathophysiological mechanism has been subject to some debate – the historical definition of tendon damage resulting from an inflammatory process alone ('tendinitis') gave way to the degenerative model ('tendinopathy') following evidence of a dearth of inflammatory cells in damaged tendons (Järvinen, Józsa et al. 1997, Riley 2008) and became the prevailing paradigm for chronic tendon injury (Rees, Stride et al. 2014). Understanding of the model is complicated by the rise in inflammatory mediators (e.g. IL-6) (Andersen, Pingel et al. 2011) and promotion of collagen synthesis in the peritendinous region in response to physiological mechanical loading and evidence of very low rates of tendon remodeling (Kjaer, Bayer et al. 2013). Recent evidence has revisited inflammation as a contributing factor to the cumulative damage model or vascular insufficiency model associated with tendinopathy, however there is still a lack of clarity. In paired samples from normal and painful Achilles tendon Pingel *et al* (2013) found that expression of inflammatory markers were significantly lower in tendinopathic regions as compared to healthy tendon after acute exercise (Pingel, Fredberg et al. 2013). Evidence from rat upper-extremity overuse models, in

contrast, demonstrated higher expression of pro-inflammatory serum cytokines and greater inflammatory changes (macrophages and CCN2/CTGF^{+} fibroblasts) within the supraspinatus tendon of aged rats compared to young controls (Kietrys, Barr-Gillespie et al. 2012). The same group, using a high repetition/low force model, demonstrated inflammatory changes within the forearm long flexor digitorum with elevated serum and tissue levels of TNF-alpha and IL-6, in addition to elevated serum levels of fibrogenic proteins, TGF-β1, CCN2/CTGF, PDGF-BB AND PDGF-AB (Gao, Fisher et al. 2013). Considering loss of function in the tendon may require reconsidering active infiltration of inflammatory cells into the tissue resembling the current understanding in osteoarthritis (Rees, Stride et al. 2014).

Summary

In summary, cartilage and tendon share a limited capacity to respond to injury with repair failing to reach functional equivalence with the uninjured state. Morbidity associated with musculoskeletal disease is significant, but knowledge of the true incidence of chronic, degenerative disorders of cartilage and tendon may be underestimated. Furthermore, time scales for development of clinical signs are often protracted complicating an understanding of the pathophysiology; in tendon pathophysiology of injury is not clear with degenerative and inflammatory components described. Factors such as IL-6 appear to contribute to both de- and re-generative phenotypes. The following section considers the current therapeutic approaches and organotypic culture models for both cartilage and tendon.

1.8: Repair & Regeneration

1.8.1: Tissue engineering

Tissue-engineering and regenerative medicine (TERM) seeks to develop and synthesise functional replacements for diseased and damaged tissue in a physiologically relevant way often by the differentiation of progenitor cells. Tissues are complex, comprising multiple cells types, which may include luminal interfaces (lung), vascular and nervous tissue components, they are often multi-functional (liver), and have specialist biomechanical properties (cartilage, tendon). Replacement is also confounded by the underlying pathophysiology. Additionally, the use of synthetic biomaterials and biologically active factors are often also elements of bioengineered tissue (Atala, Kasper et al. 2012). Tissue complexity, therefore, requires a comparable level of complexity in tissue-engineering strategies. The goals and philosophies of systems biology are well-placed to inform TERM through the modeling of cell signalling and behavioural phenotypes (Cosgrove, Griffith et al. 2008) especially as regenerative strategies often aim to recapitulate dynamic processes, e.g. tissue morphogenesis. It is proposed that systems biology approaches (data integration, data mining, machine learning, network algorithms, pathway analysis) using multi-level data sources (transcriptomics, proteomics, functional annotations, and interaction networks) can facilitate the development of predictive models for bioengineered tissues to inform and refine these systems (Rajagopalan, Kasif et al. 2013). Systems biology tools are applied in this thesis to explore novel approaches to inform TERM.

Organotypic culture systems

Tissue-specific environments are preserved contexts that confer function through biophysical, biochemical and biological signals. Homeostasis and morphogenesis entail

dynamic dialogues between multiple cell types responding to cues coded in various forms on multiple hierarchical levels and within temporal and spatial scales of several orders of magnitude (Johnson, Leight et al. 2007). The restitution of tissue function *ex vivo* has been a common goal for biologists that has led to the development of diverse methods referred to as 'three-dimensional', 'organotypic', or 'organoid' culture (Shamir and Ewald 2014). Organotypic cultures represent a sub-discipline of TERM and use bioengineering techniques to develop biomimetic environments for *in vitro* tissue models for research pursuits.

Physiological relevance of *in vitro* culture

Two dimensional, monolayer, *in vitro* culture systems have been a standard tool for biologists for over a century (Harrison, Greenman et al. 1907) facilitating manipulation of live cells *ex vivo*. Contemporaries of Harrison (1902), the histologists Golgi and Ramon y Cajal, were awarded the Nobel Prize in Physiology and Medicine for their images of neural structures (De Carlos and Borrell 2007, Musumeci 2014). Their drawings and histological images were considered representative of neural tissue organization in two-dimensions; these drawing alone were more complex renderings of the *in vivo* environment than monolayer culture. Although the simplicity of the monolayer system has aided reductionist biology it is self-evident that monolayer cultures in no way adequately recapitulate complex three-dimensional tissue architecture or the developmental or homeostatic conditions within which cells function (Baker and Chen 2012). Yet monolayer culture is still the basis of most *in vitro* work in cartilage and tendon biology.

Conventional *in vitro* studies grossly underestimate the complexity of the native environment by growing heterogenous cell populations on tissue culture plastics, under hyperoxic conditions, non-physiological concentrations of bio-active factors, and devoid

of mechanical stimulation. Promotion of a proliferative profile and metabolic alterations compound phenotypic drift and senescence - as such, given the clear inadequacies of traditional methods, the question has been posed by Spanoudes, *et al* (2014), as to how much longer 'physiologically irrelevant' culture systems may be employed (Spanoudes, Gaspar et al. 2014).

Summary

The ultimate objective of *in vitro* tissue engineering, and so also organotypic cultures, is to comprehensively, consistently and consummately replicate the *in vivo* environment in surrogate tissue cultures, in a physiologically relevant manner. Additionally, they serve as an opportunity to integrate disease models with rational therapeutic design in a dynamic setting. In reality, recreating the complex relationships in part, through a simple, cost-effective and reproducible model that does not compromise cellular behavior, would suffice (Johnson, Leight et al. 2007).

1.8.2: Obstacles to development of organotypic cultures

Progress in the development of organotypic cultures is related to difficulties in replicating the three-dimensional structures (e.g. highly axial structure in tendon), ensuring the major constituents and cell populations are in correct ratios, and harvesting sufficient number of cells for autologous therapies. The general approaches to qualifying the progress from novel interventions in organotypic cultures and regenerative therapies has been histological and biomechanical properties and targeted, yet limited, gene and protein profiling.

For clinically relevant cell-based therapies to be attained they require an integrated approach to prevent transdifferentiation of cells (towards an alternative cell phenotype) and commitment of the cell population to a stable, functional phenotype. This would

require a concerted interaction of biophysical (topographical and mechanical contexts), biochemical environment cues (oxygen tension, pH, metabolic precursors) and dynamic biological signals (growth factors, cytokines, and co-cultures) with relevant temporal relationships and kinetics (Spanoudes, Gaspar et al. 2014). As mentioned earlier in this section systems biology techniques are especially suited to this problem of complexity and integration.

Spanoudes, *et al* (2014) cite a number of obstacles to the development of adequate organotypic culture models for tendon, although they have more general application to cartilage models too: a) a lack of standardised and commercially available products (e.g. fabricated scaffolds, bioreactor systems); b) a lack of protein and gene analysis tools of sufficient depth and common potential; c) little collaborative and inter-disciplinary analysis; d) continued use of mono-dimensional culture systems (Spanoudes, Gaspar et al. 2014). Further, the authors describe that extrapolating definitive conclusions across studies is prohibited by an *ad hoc* approach to interventions (such as surface topology), a lack of agreement on differentiation markers and standardised read-outs (morphology, gene and protein expression), or uniform transparency of cell sourcing, harvesting and culture methods. In a review by Johnson, *et al* (2007), the authors note that many three-dimensional culture systems incorporate multiple modifications simultaneously, often crudely defining the ECM components (Johnson, Leight et al. 2007). This highlights the requirement for a clear understanding of simple organotypic systems rather than the development of increasingly more complex bio-engineered tissues.

1.8.3: Repairing musculoskeletal tissues

Intrinsically there is little scope for regeneration following injury, with insidious remodelling and matrix degradation prevailing (Steinert, Ghivizzani et al. 2007). Perturbations to matrix homeostasis are likely to be an early event in the onset of

osteoarthritis (Heinegard and Saxne 2011), however, pathogenesis and genetic predispositions are still subject to much discussion (Berenbaum 2013, Felson 2013, Reynard and Loughlin 2013). Tendon, for which there is a dearth of data, has similarly poor reparative responses for comparable reasons (Kannus 2000, Liu, Aschbacher-Smith et al. 2011, Young 2012). Medical interventions for the facilitation of healing and regeneration of cartilage and tendon, such as autologous cartilage implantation, or the use of bone-marrow or adipose-derived mesenchymal stem cells, to treat tendon injury inherently rely on the expansion of cells in monolayer culture prior to implantation. These early interventions often had little or no validation in terms of the alterations that such a considerable change in environment has had on the global synthetic profile of these cells (Demoor, Maneix et al. 2012).

Stem and progenitor cells

Although not the focus of this thesis considerable reference is made to stem/progenitor cells in the context of tissue engineering and regenerative therapies and so these will be reviewed in brief.

Stem cells are an autonomous population of self-renewing (clonogenic) cells with the capacity to differentiate into multiple tissue lineages in response to biological cues. To understand the motivation of cellular regenerative strategies in musculoskeletal biology the most recent developments in stem cell research should be considered. These cells hold particular value as they may be autologous and side-step the contentious use of embryonic-derived cells.

It is widely accepted that a progenitor cell population, variably termed 'adult-somatic stem cells' or 'tissue stem cells', exist within variable niches in adult tissue (Fuchs and Horsley 2011). Niches represent highly specialized environments that facilitate the

balance between self-renewal and differentiation (Moore and Lemischka 2006). These cells are distinct from embryonic or induced pluripotent cells in that they have a reduced capacity to differentiate into different tissue lineages. It is striking that progenitor cells from mesenchymal tissues (adipose, muscle, cartilage, bone) are shown reduced efficiency for multi-lineage differentiation as compared to their capacity to differentiate towards their tissue of origin (Pizzute, Lynch et al. 2015). This suggests that progenitor cells derived from adult tissue have an inherently constrained plasticity, likely to be epigenetically regulated.

Progenitor cells are often identified by a myriad of cell surface markers (CD90/THY1, CD166, SSEA-4, D146) or by gene expression (*Oct4*, *Nanog*, *Ktf4*), however, there is considerable transcriptomic and proteomic heterogeneity between studies and tissue sources and expression is not stable *in vitro* (Lv, Tuan et al. 2014).

Musculoskeletal progenitor cells

As presented in section **1.7.4** cartilage and tendon progenitors in the embryonic developing limb have been shown to express SCX and SOX9 in various combinations associated with anatomical location. This does not, however, reflect the variety of markers in recent studies defining musculoskeletal progenitors. Recently embryonic osteochondroreticular stem cells have been identified, by BMP-antagonist *Grem1* expression, that retain the capacity to differentiate into bone, cartilage and synovial cell phenotypes (Worthley, Churchill et al. 2015). In cartilage research attention has also focused on migratory progenitor cells (Koelling, Kruegel et al. 2009). In Schminke and Miosge (2014) chondrogenic progenitor cells with a migratory phenotype, under the influence of SOX9 and RUNX2, were shown to express stem cell surface markers (CD105, CD73, Stro-1) (Schminke and Miosge 2014), however this study mainly demonstrated cells with a fibroblastic phenotype arising from tissue explants *in vitro*.

Although transcriptome analysis by microarray has been undertaken for these cell populations there has been no comparison with normal tissue or other culture phenotypes. Consequently, although potential sub-populations of progenitor cells have been defined in contemporary studies they still lead by assertion with populations defined by expression of a limited panel of markers and with no comparison to *in vitro* phenotypes. In this thesis baseline gene and protein expression surveys serve to inform researchers of the synthetic profile of chondrocytes and tenocytes in different environments.

Musculoskeletal tissue regeneration strategies

Three principal strategies have been proposed to pursue regenerative potential in mammals: i) stem cell transplantation (physical implantation of cells), ii) cells seeding and embedding into scaffolds, iii) dedifferentiation of terminally differentiated cells (Odelberg 2002, Cai, Fu et al. 2007, Maden 2013). The latter has been largely overlooked as a reparative approach in cartilage and tendon and is this is addressed in this thesis, **Chapter 2**. A brief overview of alternative strategies for cartilage and tendon repair is provided below.

Cartilage repair: autologous cartilage implantation

Therapeutic interventions in OA range from arthroscopic interventions, such as micro-fracture and autologous chondrocyte implantation (ACI), scaffold-based and scaffold-free techniques (Makris, Gomoll et al. 2015) to total joint replacement. ACI requires the harvesting of normal, healthy cartilage from non-load-bearing hyaline cartilage of diseased joints to be implanted into early chondral lesions following expansion of isolated cells in monolayer culture (Brittberg, Lindahl et al. 1994). As will be discussed further in **Chapter 2** this deprives the resident cells of the complex three-dimensional environment and induces a state of dedifferentiation, in part mimicking aspects of

osteoarthritis (Demoor, Ollitrault et al. 2014). Consequently, the treated chondral defect heals as fibrocartilage. Three separate systematic Cochrane reviews (2002–2010) of ACI in the knee concluded that despite increasing and widespread use of the technique there was insufficient evidence to draw conclusions on the clinical impact (Wasiak and Villanueva 2002, Wasiak, Clar et al. 2006, Vasiliadis and Wasiak 2010). Central to this methodology is the assumption that the implanted chondrocytes retain functionality, are a homogenous population, are not senescent and will be viable when transferred from a hyperoxic culture environment into the hypoxic joint environment.

1.8.4: Parallels in tendon healing and development

The temporal expression of growth factors, cell proliferation, migration and ECM production in tendon healing parallels that of embryonic tendon development; as such, some strategies for tendon regeneration have focused on harnessing these similarities (Connizzo, Yannascoli et al. 2013). Cytokine modulators of tendon healing include: basic FGF, TGF-β, PDGF, BMP family members, IGF and VEGF (Müller, Todorov et al. 2013).

Methods for tendon repair and regeneration

Excluding direct surgical intervention, anti-inflammatory and rehabilitative therapies (shock-wave therapy, therapeutic ultrasound) there are three principal, although not mutually exclusive, strategies being employed and researched to promote tendon healing. First, regenerative strategies based upon the application of soluble cues to facilitate tenogenic differentiation; second, tissue engineering techniques, incorporating mechanical cues and environmental factors into the development of replacement tissue; thirdly, cell-based therapies (Ho, Sawadkar et al. 2014).

To develop organo-typic structures the complex architecture of tendons would have to be recreated. Principle amongst these would be: a) alignment of collagen fibres and fascicular structure; b) transitional properties, e.g. proteoglycan content, mineralization and collagen type that would be found at the myotendinous junction, bone insertion sites and areas of compression; c) the degree of collagen cross-linking, and d) the collagen fibre crimp pattern (Connizzo, Yannascoli et al. 2013). In injuries such as anterior cruciate ligament rupture and rotator cuff injury there often little in the way of useful healthy tissue available for repair (Müller, Todorov et al. 2013) further complicating repair strategies.

Soluble factors

Understanding of the role of these soluble factors in development and healing have led to the use of autologous sources of growth factors, usually in the form of platelet-rich plasma (PRP); the alpha granules of platelets contain many of the growth factors and cytokines listed earlier. Despite the attractive nature of an autologous source of growth factors, and their gaining popularity, the evidence for early intervention with PRP is not convincing. A recent well-controlled study (Kaniki, Willits et al. 2014) and systematic review (de Vos, van Veldhoven et al. 2010) have failed to demonstrate a clear benefit for the use of PRP in acute tendon injury or chronic tendinopathy. Another key growth factor undergoing active research for tendon and ligament repair is platelet-derived growth factor BB (PDGF-BB) (Thomopoulos, Das et al. 2009, Shah, Bendele et al. 2013) with which there is some evidence for improved functional repair over PRP or corticosteroids (Solchaga, Bendele et al. 2014). The relevance of PDGF BB to differentiation of tenocytes is considered further in **Chapter 7**.

Cell-therapies

A variety of cell sources have been used for cell replacement therapy studies including, bone-marrow or adipose-derived MSCs and autologous fibroblasts/tenocytes (Ho, Sawadkar et al. 2014). Tendon-derived stem cells (Bi, Ehirchiou et al. 2007) may be an alternative source of autologous cells for tendon repair (Lui and Chan 2011). The definition of the differentiated status of tenocytes varies between studies. Many studies considering the development of regenerative tendon models use a narrow definition of tendon differentiated status, for example, the expression of type-I collagen, and/or the expression of the differentiation marker TNMD. For example, Tan, *et al* (2014) recently reported improved repair of a window injury model of the rat patellar tendon using SCX-transduced tendon-derived stem cells (Tan, Lui et al. 2014). Using a fibrin scaffold containing a high density of late-passage cells fibrin constructs were placed into the injured patellar window. The authors report improved histological scoring and repair, higher expression of COL1A1 by immunohistochemistry (IHC), and higher levels of expression of tenogenic markers by qPCR. Changes between the controls and transduced constructs, however, appeared equivocal by eight weeks post-implant. In **Chapters 2** and **3** the synthetic profiles of tenocytes in fibrin cultures are defined.

Scaffolds

Artificial scaffolds have been a focus of keen research and strong commercialization. Scaffolds attempt to provide a microenvironment that may support damaged tissue biomechanically and stimulate cell proliferation and growth. As host tendon is often damaged, tendon allografts and xenografts have found popularity for the repair of large tendon and ligament defects. These may be biological, decellularised extra-cellular matrices derived from animal or human connective tissue (e.g. porcine small intestinal submucosa), or synthetic, chemically-derived products (e.g. polyester, polypropylene).

Additionally, most studies concerning commercial products are retrospective or small case studies, limiting any insight into the benefits of these products (Chen, Xu et al. 2009).

Summary

The plethora of regenerative interventions that have developed for cartilage and tendon repair suggest that few have been universally successful. Systematic evaluation of the interactions between cells, biomaterials, bio-active factors and other environmental variables is required. This development of standardised biomimetic culture systems would aid the validation of developments in tissue engineering.

1.8.5: Musculoskeletal organotypic/three-dimensional models

The provision of normal human tissue for musculoskeletal research is often a limiting factor for achieving adequately powered studies. Standardised organotypic, three-dimensional culture systems would ideally limit the requirement of healthy tissue, or *in vivo* models, for predictive and prognostic assays, and preclinical therapeutic testing for genome editing and therapeutics. The approach to three-dimensional culture development for musculoskeletal tissues has often followed a deconstructed approach, i.e. considering one tissue in isolation or using one mechanism (mechanical or biological cues) to optimize a system. Pertinent research relating to two common *in vitro* models for cartilage and tendon, the focus in this thesis, is outlined below.

Alginate-Encapsulated chondrocyte cultures

Alginate-encapsulated chondrocytes have been widely used as a three-dimensional culture system for the re-differentiation of monolayer-expanded chondrocytes, **Figure 1.1**. Alginic acid is a naturally occurring polysaccharide derived from seaweed and is especially attractive for tissue engineering as it is considered to have excellent

biocompatibility (it is an Food and Drug Administration (FDA)-approved polymer), may be processed in a number of ways, may be utilized for drug or cell delivery and may be modified chemically and physically to reflect the application (Sun and Tan 2013).

Benya and Schaffer (1982) described the use of agarose, another polysaccharide polymer, to promote a chondrogenic synthetic profile from dedifferentiated chondrocytes (Benya and Shaffer 1982). Transfer to agarose suspension cultures from monolayer and a reversion to the differentiated phenotype initially resulted in reduced rates of collagen, proteoglycan, DNA and protein synthesis, with approximately 80% of cells surviving the transition. Although DNA and protein synthesis did not reach the levels at culture initiation proteoglycan and collagen synthesis (with α1(II)-chains showing elevated levels within two days of culture) increase significantly over a period of 14 days.

Almqvist *et al* (2001) reported that chondrocytes proliferate in alginate cultures by showing out-growth of collagen type II-positive cells into a cell-free fibrin gel surrounding the alginate beads and pronounced increases in DNA and aggrecan synthesis (Almqvist, Wang et al. 2001). However, other reports contradict the proliferative capacity of chondrocytes in alginate: Müller, *et al* (2008) found introduction to alginate cultures retarded proliferative activity relative to monolayer, but also that a small, rapidly proliferating cell population was present after eight days in culture (Müller, John et al. 2008). Extended population doubling times relative to monolayer have also been presented elsewhere (Baghaban Eslaminejad, Taghiyar et al. 2009). Jonitz, *et al* (2011) found that dedifferentiated chondrocytes in unstimulated alginate cultures showed a significant reduction in DNA synthesis over 35 days in culture relative to alginate cultures in pro-chondrogenic media containing TGF-β1 and IGF-1 (Jonitz, Lochner et al. 2011). Notably, this study demonstrated the increased expression of

MSC-associated cell-surface markers (CD104, CD44, CD166 and CD29) in monolayer-expanded chondrocytes. A stimulation of re-differentiation by TGF-β1 was also shown in a study considering the effect of different sources of chondrocytes. Relative to ear and nose chondrocytes, or BM-MSCs, articular chondrocytes demonstrated the highest chondrogenic potential in alginate cultures, but this was not reproduced *in vivo* (Pleumeekers, Nimeskern et al. 2014). Hypoxia has been demonstrated to improve the redifferentiation capacity of bovine articular chondrocytes in alginate cultures with a 5% oxygen tension showing elevated collagen type II production by Western blot (Domm, Schünke et al. 2002). These studies are difficult to compare with a range of seeding densities used ($5x10^5$ – $5x10^6$/mL), culture durations (7 days-2 months), varying oxygen tensions (5-21%), chondrocyte sources (articular, nasal septum, OA-derived) and the extent of dedifferentiation (period in monolayer culture) is not standardised.

Notably, the synthetic profiles considered in the aforementioned studies have been limited. In general expression of collagen type II and aggrecan are considered, in addition to the production of GAGs. More recent studies have included more diverse consideration of synthetic profiles. Caron, *et al* (2012) established in a study comparing human articular chondrocytes (HACS) in high-density monolayer culture to alginate bead or pellet cultures, that three-dimensional (3D) cultures resulted in higher expression of *Col2a1* and *Acan* as expected, but also *Sox9* (especially in alginate beads) (Caron, Emans et al. 2012). Notably, protein levels of COL2A1 and SOX9 were not significantly different between 2D and 3D culture systems by quantitative immunoblots. Hypertrophic synthetic profiles, characterised by the higher expression of alkaline phosphatase, MMP13, COL10A1 and RUNX2, were associated with monolayer chondrocytes after 7 days in the same differentiation media. Furthermore, 3D cultures

did not support cell proliferation to the same extent as monolayer culture, as defined by DNA synthesis.

Tensional fibrin constructs associated with embryonic-like fibrillogenesis

The presence of tension-dependent actin-rich fibripositors in embryonic chick tendon cells were found in an in vitro three-dimensional cell culture model (Kapacee, Richardson et al. 2008). In this model embryonic tenocytes were seeded into fibrin gels held in culture wells, Figure 1.1. Two anchor points, consisting of suture material secured in the wells by minutien pins, served to provide tension across the self-contracting culture and allowed the formation of linear, tendon-like constructs. Further work by Bayer, et al (2010) found that mature human tendon fibroblasts were able to recapitulate embryonic fibrillogenesis, using the same culture model, as defined by the presence of membrane-enclosed fibrils in fibripositors (Bayer, Yeung et al. 2010). The authors concluded that mature tendon fibroblasts retained an intrinsic capacity for embryonic collagen fibril formation when cultured under tension, however, monolayer expansion of harvested cells up to passage five may have influenced the differentiation state of the mature cells. Similarly, mesenchymal stromal cells in comparable fibrin-constructs exhibited collagen fibril-containing fibripositors and the higher expression of phosphorylated SMAD2 under the influence of TGF-β signalling (Kapacee, Yeung et al. 2010).

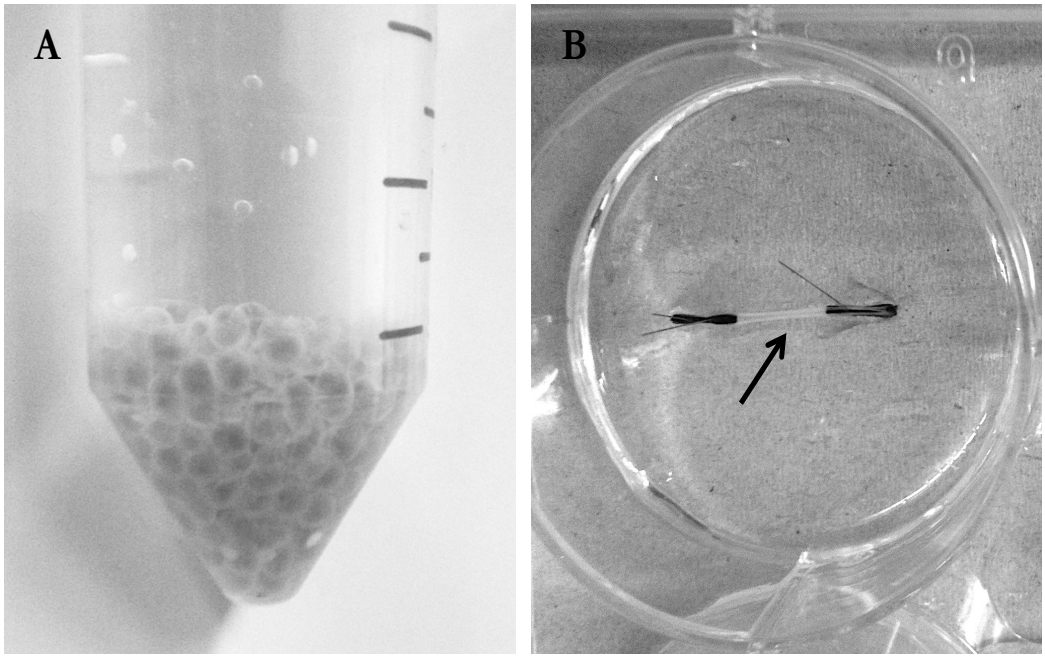


Figure 1.1: Three-dimensional culture systems are proposed to better replicate the native environment than two-dimensional monolayer culture. Two standard techniques are represented: A) chondrocytes suspended in alginate beads; B) linear tendon construct (arrow) suspended between two pieces of suture material pinned to base of six-well plate.

1.8.6: Summary statements

For both tissue engineering and organotypic culture systems the ultimate aim is the development of physiologically relevant and/or functionally equivalent proxies for the tissue under investigation/repair. Without validated *in vitro* organotypic models it is not possible to translate developments in regenerative medicine from the laboratory to the clinic in an expedient manner. Furthermore, incremental advances in the biochemical and biomechanical properties of *in vitro* systems are not made with respect to a standardised baseline system or 'worst model', rather on an individual laboratory basis. In addition, the rapid development of complex combinations of progenitor cells, biomaterials, and bio-active compounds needs to be considered with respect to the

regulatory framework for safety and efficacy in the translation of engineered tissue into humans (Atala, Kasper et al. 2012). The blind-pursuit of increasingly complex structures may ignore simple solutions, i.e. increasing complexity of a bio-engineered model does not mean that it is functionally more stable. At present organotypic models have been insufficiently described in terms of their transcriptomic and proteomic profiles, let alone bio-engineered tissue.

1.9: Closing statements

1.9.1: Project motivation

Cartilage and tendon repair strategies often have equivocal evidence for their application, with *in vitro* models and novel experimental interventions defined by a limited array of established gene expression markers or qualitative end-points. There is a need for concerted community efforts to standardise methodologies, culture systems, interrogate global gene and protein expression, and establish common functional end-points.

Understanding of regulatory networks involved in cartilage and tendon morphogenesis, homeostasis, pathology and regeneration is largely incomplete and limits attempts to develop comprehensive systems models. Deconstructed, 'bottom-up' approaches to defining these are limited in scope and potential and so integrated, holistic approaches are required if rational *in vitro* models or regenerative therapies are to be developed. The common origins, structural components, integrated development, and functionality of cartilage and tendon should be considered when organotypic systems are developed. Establishing regulatory networks, through gene expression and protein abundance

profiling, which may influence the differentiation status of both cartilage and tendon would provide an evidence-based rational for developments in tissue engineering.

A systems biology approach lends itself to this 'top-down' integration of multi-level data and this thesis employs such a methodology to define inform future tissue engineering strategies through the exploration of organotypic, three-dimensional culture systems. The genome may be considered a database of coded functions rather than a formal program for life. This database of functions or subroutines (proliferation, apoptosis, etc) may be accessed in a cell-autonomous manner as required and used many times over (Bard 2013). Whilst some of these subroutines are well-described de- and re-differentiation in the adult mammalian cell have received little attention. The potential for dedifferentiation as a regenerative mechanism has been alluded to in this introduction. Using systems biology approaches it is possible to define highly co-expressed sub-networks of genes with strong phenotypic associations that may reveal gene regulatory networks relevant to these processes.

1.9.2: Project outline

In this thesis the system under consideration are the major cell populations of cartilage tendon (chondrocytes and tenocytes, respectively) and they are described in three environmental conditions: native, monolayer (two-dimensional), or three-dimensional models, **Figure 1.2**. There has been no systematic investigate of the global gene and protein profiles of cartilage and tendon in their native state relative to monolayer or three-dimensional cultures. There is no clear mechanistic description of the impact that *in vitro* environmental perturbations have on the system or indeed the adequacy of these models as proxies for cartilage and tendon. Monolayer represents the most common *in vitro* model yet there is little evidence for validation of growth of cartilage or tendon cells in culture relative to their native tissues or model organotypic cultures.

A discovery approach using transcriptomic and proteomic profiling is undertaken in the first instance to define a robust and consistent gene and protein profile for each condition. Differentially expressed elements are functionally annotated and pathway topology approaches employed to predict major signalling pathways associated with the observed phenotype

Gene-gene co-expression network analysis is being increasingly used and a systems biology method to extract molecular networks from multi-dimensional data and describe multiple regulatory systems activities. They are also being used to generate novel hypotheses about complex disease mechanisms as they are not constrained by prior knowledge of molecular biology (Gaiteri, Ding et al. 2014). This methodology is most commonly used to define gene 'hubs' with strong disease correlation, which may represent putative biomarkers. However, this technique shows greater potential when employed to extract co-expression networks, which represent the underlying descriptors of the system and may be used to translate multi-scale data sets into testable predictions.

In this thesis gene-gene co-expression network analysis is employed for transcriptome network decomposition to isolate highly correlated and interconnected gene-sets (modules) from gene expression profiles of cartilage and tendon cells in different environmental conditions in order to define sub-networks regulating de- and re-differentiation. Gene expression data from discovery studies and publically available transcriptomic data sets are integrated as part of a weighted gene co-expression network analysis (WGCNA). Comparison of global transcriptome network architecture is performed to define the conservation of network modules between a model species (rat) and human data. Finally, initial findings from the integration of gene and protein profiles are presented and mechanistic networks derived to described de- and re-differentiation in chondrocytes and tenocytes.

The studies collected in this thesis contribute to a wider understanding of cartilage and tendon tissue engineering and organotypic culture development. A clear mechanistic understanding of the regulatory networks controlling differentiation of cartilage and tendon progenitor cells is required in order to develop improved *in vitro* models and bio-engineered tissue that are physiologically relevant.

1.9.3: Project objectives

This project was devised and executed with the intention of tackling one aspect of the present limitations of tissue engineering for cartilage and tendon, i.e. the dearth of in-depth transcriptomic and proteomic profiling and integrate data to explore emerging relationships that may describe these organotypic systems. Beyond the generation of reference data sets the project aimed to deliver three key objectives:

- Objective 1: To define dedifferentiation and re-differentiation in terms of synthetic profile and mark-out the phenotypic boundaries within which cartilage and tendon cells function;
- Objective 2: Define cross-species responses to homeostatic perturbations in cartilage and tendon through integration of gene-expression data from multiple gene expression studies;
- Objective 3: Integrate gene expression and protein abundance data to rationalize validation targets and derive mechanistic networks.

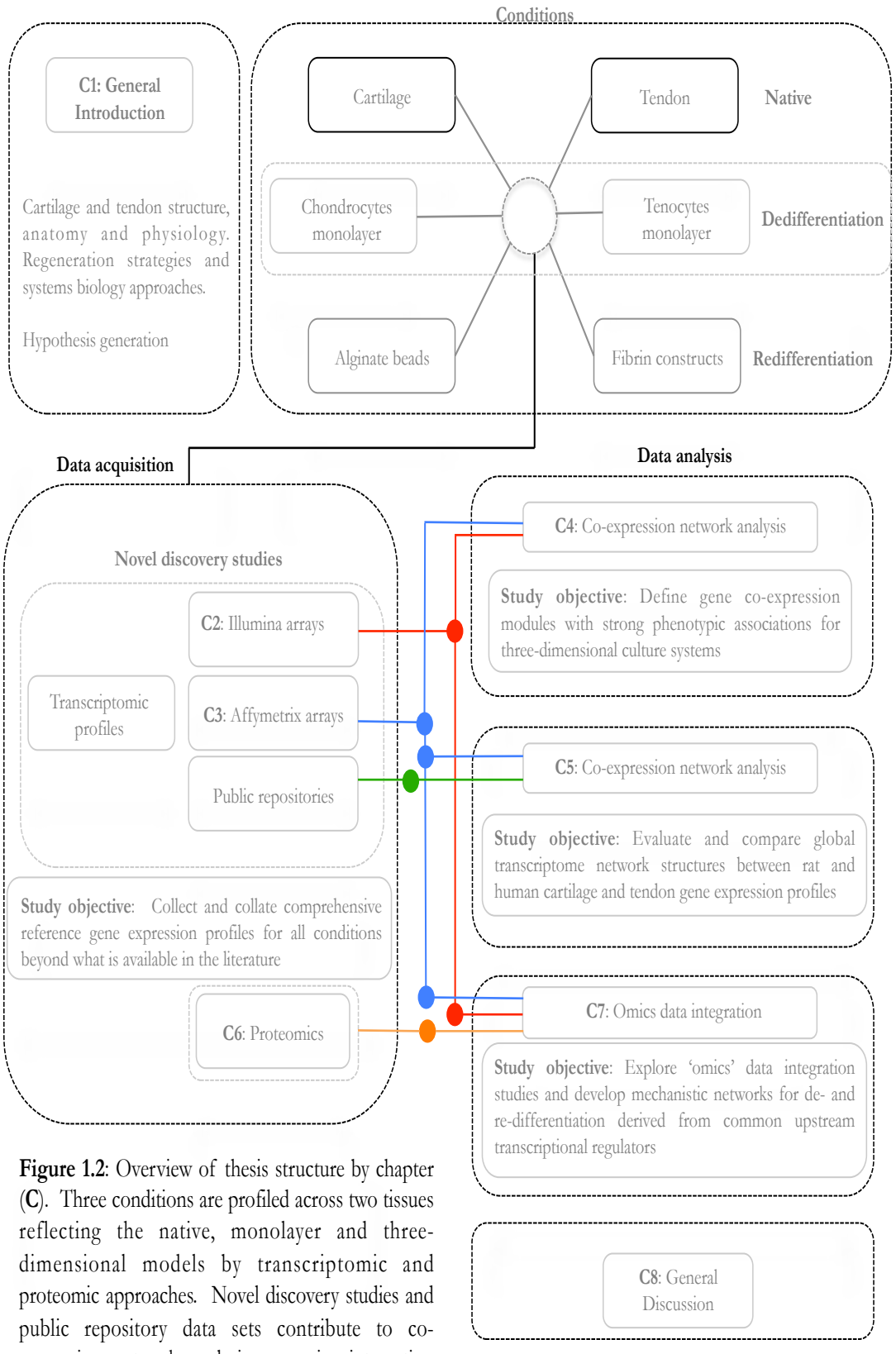


Figure 1.2: Overview of thesis structure by chapter (**C**). Three conditions are profiled across two tissues reflecting the native, monolayer and three-dimensional models by transcriptomic and proteomic approaches. Novel discovery studies and public repository data sets contribute to co-expression network analysis or omics integration analysis. Major study objectives are outlined with regard to the hypotheses outlined in the introduction and subsequent chapters. A general discussion unifies findings from all studies.

References

Almonte-Becerril, M., M. Costell and J. B. Kouri (2014). "Changes in the integrins expression are related with the osteoarthritis severity in an experimental animal model in rats." Journal of Orthopaedic Research **32**(9): 1161-1166.

Almqvist, K. F., L. Wang, J. Wang, D. Baeten, M. Cornelissen, R. Verdonk, E. M. Veys and G. Verbruggen (2001). "Culture of chondrocytes in alginate surrounded by fibrin gel: characteristics of the cells over a period of eight weeks." Annals of the Rheumatic Diseases **60**(8): 781-790.

Andersen, M. B., J. Pingel, M. Kjær and H. Langberg (2011). "Interleukin-6: a growth factor stimulating collagen synthesis in human tendon." Journal of Applied Physiology (Bethesda, Md. : 1985) **110**(6): 1549-1554.

Anderson, D., J. Arredondo, K. Hahn, G. Valente, J. Martin, J. Wilson-Rawls and A. Rawls (2006). "Mohawk is a novel homeobox gene expressed in the developing mouse embryo." Dev. Dyn. **235**(3): 792-801.

Anderson, D., B. Beres, J. Wilson-Rawls and A. Rawls (2009). "The homeobox gene Mohawk represses transcription by recruiting the sin3A/HDAC co-repressor complex." Dev. Dyn. **238**(3): 572-580.

Aslan, H., N. Kimelman-Bleich, G. Pelled and D. Gazit (2008). "Molecular targets for tendon neoformation." The Journal of Clinical Investigation **118**(2): 439-444.

Asou, Y., A. Nifuji, K. Tsuji, K. Shinomiya, E. Olson, P. Koopman and M. Noda (2002). "Coordinated expression of scleraxis and Sox9 genes during embryonic development of tendons and cartilage." Journal of Orthopaedic Research **20**(4): 827-833.

Atala, A., K. Kasper and A. Mikos (2012). "Engineering Complex Tissues." Science Translational Medicine **4**(160): 160rv112-160rv112.

Baghaban Eslaminejad, M., L. Taghiyar and F. Falahi (2009). "Quantitative analysis of the proliferation and differentiation of rat articular chondrocytes in alginate 3D culture." Iranian Biomedical Journal **13**(3): 153-160.

Baker, B. and C. Chen (2012). "Deconstructing the third dimension: how 3D culture microenvironments alter cellular cues." Journal of Cell Science **125**(Pt 13): 3015-3024.

Balbin, A., J. Prensner, A. Sahu, A. Yocum, S. Shankar, R. Malik, D. Fermin, S. Dhanasekaran, B. Chandler, D. Thomas, D. Beer, X. Cao, A. Nesvizhskii and A. Chinnaiyan (2013). "Reconstructing targetable pathways in lung cancer by integrating diverse omics data." Nature Communications **4**: 2617.

Banos, C., A. Thomas and C. Kuo (2008). "Collagen fibrillogenesis in tendon development: current models and regulation of fibril assembly." Birth Defects Research. Part C, Embryo Today : Reviews **84**(3): 228-244.

Bard, J. (2013). "Systems biology - the broader perspective." Cells **2**(2): 414-431.

Bayer, M., C.-Y. Yeung, K. Kadler, K. Qvortrup, K. Baar, R. Svensson, P. Magnusson, M. Krogsgaard, M. Koch and M. Kjaer (2010). "The initiation of embryonic-like collagen fibrillogenesis by adult human tendon fibroblasts when cultured under tension." Biomaterials **31**(18): 4889-4897.

Benjamin, M., E. Kaiser and S. Milz (2008). "Structure-function relationships in tendons: a review." Journal of Anatomy **212**(3): 211-228.

Benya, P. and J. D. Shaffer (1982). "Dedifferentiated chondrocytes reexpress the differentiated collagen phenotype when cultured in agarose gels." Cell **30**(1): 215-224.

Berenbaum, F. (2013). "Osteoarthritis as an inflammatory disease (osteoarthritis is not osteoarthrosis!)." Osteoarthritis and Cartilage **21**(1): 16-21.

Bhosale, A. and J. Richardson (2008). "Articular cartilage: structure, injuries and review of management." British Medical Bulletin **87**(1): 77-95.

Bi, Y., D. Ehirchiou, T. Kilts, C. Inkson, M. Embree, W. Sonoyama, L. Li, A. Leet, B.-M. Seo, L. Zhang, S. Shi and M. Young (2007). "Identification of tendon

stem/progenitor cells and the role of the extracellular matrix in their niche." Nature Medicine **13**(10): 1219-1227.

Blitz, E., A. Sharir, H. Akiyama and E. Zelzer (2013). "Tendon-bone attachment unit is formed modularly by a distinct pool of Scx- and Sox9-positive progenitors." Development (Cambridge, England) **140**(13): 2680-2690.

Bourne, A., S. Whittle, B. Richards, C. Maher and R. Buchbinder (2014). "The scope, funding and publication of musculoskeletal clinical trials performed in Australia." The Medical Journal of Australia **200**(2): 88-91.

Brent, A. E., R. Schweitzer and C. J. Tabin (2003). "A Somitic Compartment of Tendon Progenitors." Cell **113**(2): 235-248.

Breslauer, D., P. Lee and L. Lee (2006). "Microfluidics-based systems biology." Molecular bioSystems **2**(2): 97-112.

Brittberg, M., A. Lindahl, A. Nilsson, C. Ohlsson, O. Isaksson and L. Peterson (1994). "Treatment of Deep Cartilage Defects in the Knee with Autologous Chondrocyte Transplantation." N Engl J Med **331**(14): 889-895.

Brown, J., V. Finley and C. Kuo (2014). "Embryonic mechanical and soluble cues regulate tendon progenitor cell gene expression as a function of developmental stage and anatomical origin." Journal of Biomechanics **47**(1): 214-222.

Cai, S., X. Fu and Z. Sheng (2007). "Dedifferentiation: A New Approach in Stem Cell Research." BioScience **57**(8): 655-662.

Canty, E. and K. Kadler (2005). "Procollagen trafficking, processing and fibrillogenesis." Journal of Cell Science **118**(Pt 7): 1341-1353.

Caron, M. M., P. J. Emans, M. M. Coolsen, L. Voss, D. A. Surtel, A. Cremers, L. W. van Rhijn and T. J. Welting (2012). "Redifferentiation of dedifferentiated human articular chondrocytes: comparison of 2D and 3D cultures." Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society **20**(10): 1170-1178.

Chang, S.-W., S. Shefelbine and M. Buehler (2012). "Structural and Mechanical Differences between Collagen Homo- and Heterotrimers: Relevance for the Molecular Origin of Brittle Bone Disease." Biophysical Journal **102**(3): 640-648.

Chen, J., J. Xu, A. Wang and M. Zheng (2009). "Scaffolds for tendon and ligament repair: review of the efficacy of commercial products." Expert Review of Medical Devices **6**(1): 61-73.

Christley, S., M. Alber and S. Newman (2007). "Patterns of mesenchymal condensation in a multiscale, discrete stochastic model." PLoS Computational Biology **3**(4).

Cicuttini, F. and A. Wluka (2014). "Osteoarthritis: Is OA a mechanical or systemic disease?" Nat Rev Rheumatol **10**(9): 515-516.

Clayton, R. and C. Court-Brown (2008). "The epidemiology of musculoskeletal tendinous and ligamentous injuries." Injury **39**(12): 1338-1344.

Conesa, A. and A. Mortazavi (2014). "The common ground of genomics and systems biology." BMC Systems Biology **8 Suppl 2**: S1.

Connizzo, B., S. Yannascoli and L. Soslowsky (2013). "Structure-function relationships of postnatal tendon development: a parallel to healing." Matrix Biology : Journal of the International Society for Matrix Biology **32**(2): 106-116.

Cosgrove, B. D., L. G. Griffith and D. A. Lauffenburger (2008). "Fusing Tissue Engineering and Systems Biology Toward Fulfilling Their Promise." Cellular and Molecular Bioengineering **1**(1): 33-41.

Cross, M., E. Smith, D. Hoy, S. Nolte, I. Ackerman, M. Fransen, L. Bridgett, S. Williams, F. Guillemin, C. Hill, L. Laslett, G. Jones, F. Cicuttini, R. Osborne, T. Vos, R. Buchbinder, A. Woolf and L. March (2014). "The global burden of hip and knee osteoarthritis: estimates from the Global Burden of Disease 2010 study." Annals of the Rheumatic Diseases **73**(7): 1323-1330.

Cserjesi, P., D. Brown, K. L. Ligon, G. E. Lyons, N. G. Copeland, D. J. Gilbert, N. A. Jenkins and E. N. Olson (1995). "Scleraxis: a basic helix-loop-helix protein that prefigures skeletal formation during mouse embryogenesis." Development (Cambridge, England) **121**(4): 1099-1110.

Day, T., X. Guo, L. Garrett-Beal and Y. Yang (2005). "Wnt/beta-catenin signaling in mesenchymal progenitors controls osteoblast and chondrocyte differentiation during vertebrate skeletogenesis." Developmental Cell **8**(5): 739-750.

De Carlos, J. and J. Borrell (2007). "A historical reflection of the contributions of Cajal and Golgi to the foundations of neuroscience." Brain Research Reviews **55**(1): 8-16.

de Vos, R. J., P. L. van Veldhoven, M. H. Moen, A. Weir, J. L. Tol and N. Maffulli (2010). "Autologous growth factor injections in chronic tendinopathy: a systematic review." British Medical Bulletin **95**: 63-77.

Decker, R., E. Koyama and M. Pacifici (2014). "Genesis and morphogenesis of limb synovial joints and articular cartilage." Matrix Biology : Journal of the International Society for Matrix Biology **39**: 5-10.

Dekanty, A. and M. Milan (2011). "The interplay between morphogens and tissue growth." EMBO reports **12**(10): 1003-1010.

Demoor, M., L. Maneix, D. Ollitrault, F. Legendre, E. Duval, S. Claus, F. Mallein-Gerin, S. Moslemi, K. Boumediene and P. Galera (2012). "Deciphering chondrocyte behaviour in matrix-induced autologous chondrocyte implantation to undergo accurate cartilage repair with hyaline matrix." Pathologie Biologie **60**(3): 199-207.

Demoor, M., D. Ollitrault, T. Gomez-Leduc, M. Bouyoucef, M. Hervieu, H. Fabre, J. Lafont, J.-M. Denoix, F. Audigié, F. Mallein-Gerin, F. Legendre and P. Galera (2014). "Cartilage tissue engineering: molecular control of chondrocyte differentiation for proper cartilage matrix reconstruction." Biochimica et Biophysica Acta **1840**(8): 2414-2440.

Domm, C., M. Schünke, K. Christesen and B. Kurz (2002). "Redifferentiation of dedifferentiated bovine articular chondrocytes in alginate culture under low oxygen tension." Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society **10**(1): 13-22.

Eyre, D. (2002). "Articular cartilage and changes in Arthritis: Collagen of articular cartilage." Arthritis Res **4**(1): 30-35.

Felson, D. T. (2013). "Osteoarthritis as a disease of mechanics." Osteoarthritis and Cartilage **21**(1): 10-15.

Frantz, C., K. Stewart and V. Weaver (2010). "The extracellular matrix at a glance." Journal of Cell Science **123**(Pt 24): 4195-4200.

Fuchs, E. and V. Horsley (2011). "Ferreting out stem cells from their niches." Nat Cell Biol **13**(5): 513-518.

Gaiteri, C., Y. Ding, B. French, G. C. Tseng and E. Sibille (2014). "Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders." Genes, Brain and Behavior **13**(1): 13-24.

Gao, H., P. Fisher, A. Lambi, C. Wade, A. Barr-Gillespie, S. Popoff and M. Barbe (2013). "Increased Serum and Musculotendinous Fibrogenic Proteins following Persistent Low-Grade Inflammation in a Rat Model of Long-Term Upper Extremity Overuse." PLoS ONE **8**(8): e71875.

Goldring, M. (2012). "Chondrogenesis, chondrocyte differentiation, and articular cartilage metabolism in health and osteoarthritis." Therapeutic Advances in Musculoskeletal Disease **4**(4): 269-285.

Gordon, M. and R. Hahn (2010). "Collagens." Cell and Tissue Research **339**(1): 247-257.

Guerquin, M.-J., B. Charvet, G. Nourissat, E. Havis, O. Ronsin, M.-A. Bonnin, M. Ruggiu, I. Olivera-Martinez, N. Robert, Y. Lu, K. Kadler, T. Baumberger, L. Doursounian, F. Berenbaum and D. Duprez (2013). "Transcription factor EGR1 directs tendon differentiation and promotes tendon repair." The Journal of Clinical Investigation **123**(8): 3564-3576.

Halper, J. and M. Kjaer (2014). "Basic components of connective tissues and extracellular matrix: elastin, fibrillin, fibulins, fibrinogen, fibronectin, laminin, tenascins and thrombospondins." Advances in Experimental Medicine and Biology **802**: 31-47.

Harrison, R., M. J. Greenman, F. Mall and C. M. Jackson (1907). "Observations of the living developing nerve fiber." Anat. Rec. **1**(5): 116-128.

Hasson, P. (2011). ""Soft" tissue patterning: muscles and tendons of the limb take their form." Developmental Dynamics : an official publication of the American Association of Anatomists **240**(5): 1100-1107.

Hecht, J., E. Hayes, R. Haynes and W. Cole (2005). "COMP mutations, chondrocyte function and cartilage matrix." Matrix Biology **23**(8): 525-533.

Heinegard, D. and T. Saxne (2011). "The role of the cartilage matrix in osteoarthritis." Nat Rev Rheumatol **7**(1): 50-56.

Heino, J. (2014). "Cellular signaling by collagen-binding integrins." Advances in Experimental Medicine and Biology **819**: 143-155.

Hill, T., D. Später, M. Taketo, W. Birchmeier and C. Hartmann (2005). "Canonical Wnt/beta-catenin signaling prevents osteoblasts from differentiating into chondrocytes." Developmental Cell **8**(5): 727-738.

Ho, J. O., P. Sawadkar and V. Mudera (2014). "A review on the use of cell therapy in the treatment of tendon disease and injuries." Journal of Tissue Engineering **5**: 2041731414549678.

Hunter, W. (1743). "Of the Structure and Disease of Articulating Cartilages." Philosophical Transactions of the Royal Society **42**(B): 514-521.

Ideker, T., T. Galitski and L. Hood (2001). "A new approach to decoding life: systems biology." Annual Review of Genomics and Human Genetics **2**(1): 343-372.

Järvinen, M., L. Józsa, P. Kannus, T. L. Järvinen, M. Kvist and W. Leadbetter (1997). "Histopathological findings in chronic tendon disorders." Scandinavian Journal of Medicine & Science in Sports **7**(2): 86-95.

Johnson, K. R., J. L. Leight and V. M. Weaver (2007). Demystifying the Effects of a Three - Dimensional Microenvironment in Tissue Morphogenesis. Methods in Cell Biology. W. Yu - Li and E. D. Dennis, Academic Press. **83:** 547-583.

Jonitz, A., K. Lochner, K. Peters, A. Salamon, J. Pasold, B. Mueller-Hilke, D. Hansmann and R. Bader (2011). "Differentiation capacity of human chondrocytes embedded in alginate matrix." Connective Tissue Research **52**(6): 503-511.

Kadler, K., C. Baldock, J. Bella and R. Boot-Handford (2007). "Collagens at a glance." Journal of Cell Science **120**(Pt 12): 1955-1958.

Kaniki, N., K. Willits, N. Mohtadi, V. Fung and D. Bryant (2014). "A retrospective comparative study with historical control to determine the effectiveness of platelet-rich plasma as part of nonoperative treatment of acute achilles tendon rupture." Arthroscopy : The Journal of Arthroscopic & Related Surgery **30**(9): 1139-1145.

Kannus, P. (2000). "Structure of the tendon connective tissue." Scandinavian Journal of Medicine & Science in Sports **10**(6): 312-320.

Kapacee, Z., S. Richardson, Y. Lu, T. Starborg, D. Holmes, K. Baar and K. Kadler (2008). "Tension is required for fibripositor formation." Matrix Biology **27**(4): 371-375.

Kapacee, Z., C.-Y. Yeung, Y. Lu, D. Crabtree, D. Holmes and K. Kadler (2010). "Synthesis of embryonic tendon-like tissue by human marrow stromal/mesenchymal stem cells requires a three-dimensional environment and transforming growth factor β 3." Matrix Biology **29**(8): 668-677.

Kardon, G. (1998). "Muscle and tendon morphogenesis in the avian hind limb." Development (Cambridge, England) **125**(20): 4019-4032.

Kastelic, J., A. Galeski and E. Baer (1978). "The multicomposite structure of tendon." Connective Tissue Research **6**(1): 11-23.

Khatri, P., M. Sirota and A. Butte (2012). "Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges." PLoS Comput Biol **8**(2): e1002375.

Kietrys, D., A. Barr-Gillespie, M. Amin, C. Wade, S. Popoff and M. Barbe (2012). "Aging Contributes to Inflammation in Upper Extremity Tendons and Declines in Forelimb Agility in a Rat Model of Upper Extremity Overuse." PLoS ONE **7**(10): e46954.

Kitano, H. (2002). "Systems Biology: A Brief Overview." Science **295**(5560): 1662-1664.

Kjaer, M., M. Bayer, P. Eliasson and K. Heinemeier (2013). "What is the impact of inflammation on the critical interplay between mechanical signaling and biochemical changes in tendon matrix?" Journal of Applied Physiology **115**(6): 879-883.

Klipp, E., W. Liebermeister, C. Wierling, A. Kowald, H. Lehrach and R. Herwig (2009). Systems Biology. Weinheim, Germany, Wiley-VCH Verlag GmbH & Co.

Koelling, S., J. Kruegel, M. Irmer, J. R. Path, B. Sadowski, X. Miro and N. Miosge (2009). "Migratory Chondrogenic Progenitor Cells from Repair Tissue during the Later Stages of Human Osteoarthritis." Cell Stem Cell **4**(4): 324-335.

Kuhn, T. S. (2012). The structures of scientific revolutions, The University of Chicago Press.

Langfelder, P. and S. Horvath (2008). "WGCNA: an R package for weighted correlation network analysis." BMC Bioinformatics **9**(1): 559.

Leask, A. and D. Abraham (2006). "All in the CCN family: essential matricellular signaling modulators emerge from the bunker." Journal of Cell Science **119**(Pt 23): 4803-4810.

Lefebvre, V., R. R. Behringer and B. de Crombrugghe (2001). "L-Sox5, Sox6 and Sox9 control essential steps of the chondrocyte differentiation pathway." Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society **9 Suppl A**.

Lefebvre, V., P. Li and B. de Crombrugghe (1998). "A new long form of Sox5 (L-Sox5), Sox6 and Sox9 are coexpressed in chondrogenesis and cooperatively activate the type II collagen gene." The EMBO Journal **17**(19): 5718-5733.

Lejard, V., F. Blais, M.-J. Guerquin, A. Bonnet, M.-A. Bonnin, E. Havis, M. Malbouyres, C. B. Bidaud, G. Maro, P. Gilardi-Hebenstreit, J. Rossert, F. Ruggiero and D. Duprez (2011). "EGR1 and EGR2 involvement in vertebrate tendon differentiation." The Journal of Biological Chemistry **286**(7): 5855-5867.

Liu, C.-F., L. Aschbacher-Smith, N. Barthelery, N. Dyment, D. Butler and C. Wylie (2011). "What we should know before using tissue engineering techniques to repair injured tendons: a developmental biology perspective." Tissue engineering. Part B, Reviews **17**(3): 165-176.

Liu, H., S. Zhu, C. Zhang, P. Lu, J. Hu, Z. Yin, Y. Ma, X. Chen and H. OuYang (2014). "Crucial transcription factors in tendon development and differentiation: their potential for tendon regeneration." Cell and Tissue Research **356**(2): 287-298.

Loeser, R. F. (2014). "Integrins and chondrocyte–matrix interactions in articular cartilage." Matrix Biology **39**: 11-16.

Lorda-Diez, C., J. Montero, J. Garcia-Porrero and J. Hurle (2014). "Divergent differentiation of skeletal progenitors into cartilage and tendon: lessons from the embryonic limb." ACS Chemical Biology **9**(1): 72-79.

Lui, P. P. Y. and K. M. Chan (2011). "Tendon-Derived Stem Cells (TDSCs): From Basic Science to Potential Roles in Tendon Pathology and Tissue Engineering Applications." Stem Cell Reviews **7**(4): 883-897.

Lv, F.-J., R. Tuan, K. Cheung and V. Leung (2014). "Concise review: the surface markers and identity of human mesenchymal stem cells." Stem Cells (Dayton, Ohio) **32**(6): 1408-1419.

Maden, M. (2013). "Who needs stem cells if you can dedifferentiate?" Cell Stem Cell **13**(6): 640-641.

Makris, E., A. Gomoll, K. Malizos, J. Hu and K. Athanasiou (2015). "Repair and tissue engineering techniques for articular cartilage." Nature Reviews. Rheumatology **11**(1): 21-34.

Marcum, J. A. (2009). The conceptual foundations of systems biology: An introduction, Nova Science Publishers, Inc.

Mitrea, C., Z. Taghavi, B. Bokanizad, S. Hanoudi, R. Tagett, M. Donato, C. Voichiţa and S. Drăghici (2013). "Methods and approaches in the topology-based analysis of biological pathways." Frontiers in Physiology **4**: 278.

Miura, T. and K. Shiota (2000). "TGFβ2 acts as an "Activator" molecule in reaction-diffusion model and is involved in cell sorting phenomenon in mouse limb micromass culture." Dev. Dyn. **217**(3): 241-249.

Moore, K. and I. Lemischka (2006). "Stem cells and their niches." Science (New York, N.Y.) **311**(5769): 1880-1885.

Mort, J. and C. Billington (2001). "Articular cartilage and changes in Arthritis: Matrix degradation." Arthritis Res **3**(6): 337-341.

Müller, R. D., T. John, B. Kohl, A. Feldner, H. Zreiqat, W. Ertel, D. LaFace, B. Hutchins, A. Oberholzer and G. Schulze-Tanzil (2008). "Cartilage-Specific Matrix and Integrin Expression in Three-Dimensional Articular Chondrocyte Cultures Overexpressing Human Interleukin-10." Clinical Medicine Insights: Arthritis and Musculoskeletal Disorders **1**: 21-32.

Müller, S., A. Todorov, P. Heisterbach, I. Martin and M. Majewski (2013). "Tendon healing: an overview of physiology, biology, and pathology of tendon healing and systematic review of state of the art in tendon bioengineering." Knee Surgery, Sports Traumatology, Arthroscopy.

Murray, C., T. Vos, R. Lozano, M. Naghavi, A. Flaxman, C. Michaud, ..., M. AlMazroa and Z. Memish (2012). "Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010." Lancet **380**(9859): 2197-2223.

Musumeci, G. (2014). "Past, present and future: overview on histology and histopathology." Journal of Histology and Histopathology **1**(1).

Nance, T., K. Smith, V. Anaya, R. Richardson, L. Ho, M. Pala, S. Mostafavi, A. Battle, C. Feghali-Bostwick, G. Rosen and S. Montgomery (2014). "Transcriptome analysis reveals differential splicing events in IPF lung tissue." PloS ONE **9**(3): e92111.

Nouailles, G., A. Dorhoi, M. Koch, J. Zerrahn, J. Weiner, K. Faé, F. Arrey, S. Kuhlmann, S. Bandermann, D. Loewe, H.-J. Mollenkopf, A. Vogelzang, C. Meyer-Schwesinger, H.-W. Mittrücker, G. McEwen and S. Kaufmann (2014). "CXCL5-secreting pulmonary epithelial cells drive destructive neutrophilic inflammation in tuberculosis." The Journal of Clinical Investigation **124**(3): 1268-1282.

Odelberg, S. (2002). "Inducing cellular dedifferentiation: a potential method for enhancing endogenous regeneration in mammals." Seminars in Cell & Developmental Biology **13**(5): 335-343.

Panoutsopoulou, K. and E. Zeggini (2013). "Advances in osteoarthritis genetics." Journal of Medical Genetics **50**(11): 715-724.

Pingel, J., U. Fredberg, L. R. Mikkelsen, P. Schjerling, K. M. Heinemeier, M. Kjaer, A. Harisson and H. Langberg (2013). "No inflammatory gene-expression response to acute

exercise in human Achilles tendinopathy." European Journal of Applied Physiology **113**(8): 2101-2109.

Pizette, S. and L. Niswander (2000). "BMPs are required at two steps of limb chondrogenesis: formation of prechondrogenic condensations and their differentiation into chondrocytes." Developmental Biology **219**(2): 237-249.

Pizzute, T., K. Lynch and M. Pei (2015). "Impact of tissue-specific stem cells on lineage-specific differentiation: a focus on the musculoskeletal system." Stem Cell Reviews **11**(1): 119-132.

Pleumeekers, M. M., L. Nimeskern, W. L. Koevoet, N. Kops, R. M. Poublon, K. S. Stok and G. J. van Osch (2014). "The in vitro and in vivo capacity of culture-expanded human cells from several sources encapsulated in alginate to form cartilage." European Cells & Materials **27**: 264-280.

Rajagopalan, P., S. Kasif and T. M. Murali (2013). "Systems Biology Characterization of Engineered Tissues." Annual Review of Biomedical Engineering **15**(1): 55-70.

Rankin, K. S., A. P. Sprowson, I. McNamara, T. Akiyama, R. Buchbinder, M. L. Costa, S. Rasmussen, S. S. Nathan, S. Kumta and A. Rangan (2014). "The orthopaedic research scene and strategies to improve it." The Bone & Joint Journal **96-B**(12): 1578-1585.

Rees, J., M. Stride and A. Scott (2014). "Tendons - time to revisit inflammation." British Journal of Sports Medicine **48**(21): 1553-1557.

Reynard, L. and J. Loughlin (2013). "The genetics and functional analysis of primary osteoarthritis susceptibility." Expert Reviews in Molecular Medicine **15**: e2.

Ricard-Blum, S. (2011). "The collagen family." Cold Spring Harbor Perspectives in Biology **3**(1).

Riley, G. (2008). "Tendinopathy--from basic science to treatment." Nature Clinical Practice. Rheumatology **4**(2): 82-89.

Roughley, P. (2001). "Articular cartilage and changes in Arthritis: Noncollagenous proteins and proteoglycans in the extracellular matrix of cartilage." Arthritis Res **3**(6): 342-347.

Sakabe, T. and T. Sakai (2011). "Musculoskeletal diseases—tendon." British Medical Bulletin **99**(1): 211-225.

Schaefer, L. and R. Schaefer (2010). "Proteoglycans: from structural compounds to signaling molecules." Cell and Tissue Research **339**(1): 237-246.

Schiele, N., J. Marturano and C. Kuo (2013). "Mechanical factors in embryonic tendon development: potential cues for stem cell tenogenesis." Current Opinion in Biotechnology **24**(5): 834-840.

Schminke, B. and N. Miosge (2014). "Cartilage repair in vivo: the role of migratory progenitor cells." Current Rheumatology Reports **16**(11).

Schweitzer, R., J. Chyung, L. Murtaugh, A. Brent, V. Rosen, E. Olson, A. Lassar and C. Tabin (2001). "Analysis of the tendon cell fate using Scleraxis, a specific marker for tendons and ligaments." Development **128**(19): 3855-3866.

Schweitzer, R., E. Zelzer and T. Volk (2010). "Connecting muscles to tendons: tendons and musculoskeletal development in flies and vertebrates." Development (Cambridge, England) **137**(17): 2807-2817.

Scott, J. E. (2003). "Elasticity in extracellular matrix 'shape modules' of tendon, cartilage, etc. A sliding proteoglycan-filament model." The Journal of Physiology **553**(Pt 2): 335-343.

Shah, V., A. Bendele, J. Dines, H. Kestler, J. Hollinger, N. Chahine and C. Hee (2013). "Dose-response effect of an intra-tendon application of recombinant human platelet-derived growth factor-BB (rhPDGF-BB) in a rat Achilles tendinopathy model." Journal of Orthopaedic Research **31**(3): 413-420.

Shamir, E. and A. Ewald (2014). "Three-dimensional organotypic culture: experimental models of mammalian biology and disease." Nature Reviews. Molecular Cell Biology **15**(10): 647-664.

Siebens, H. (2007). "Musculoskeletal problems as comorbidities." American Journal of Physical Medicine & Rehabilitation / Association of Academic Physiatrists **86**(1 Suppl).

Smith, E., D. Hoy, M. Cross, T. Vos, M. Naghavi, R. Buchbinder, A. Woolf and L. March (2014). "The global burden of other musculoskeletal disorders: estimates from the Global Burden of Disease 2010 study." Annals of the Rheumatic Diseases **73**(8): 1462-1469.

Smith, R. K., H. L. Birch, S. Goodman, D. Heinegård and A. E. Goodship (2002). "The influence of ageing and exercise on tendon growth and degeneration--hypotheses for the initiation and prevention of strain-induced tendinopathies." Comparative Biochemistry and Physiology. Part A, Molecular & Integrative Physiology **133**(4): 1039-1050.

Soeda, T., J. M. Deng, B. de Crombrugghe, R. Behringer, T. Nakamura and H. Akiyama (2010). "Sox9-expressing precursors are the cellular origin of the cruciate ligament of the knee joint and the limb tendons." Genesis (New York, N.Y. : 2000) **48**(11): 635-644.

Solchaga, L., A. Bendele, V. Shah, L. Snel, H. Kestler, J. Dines and C. Hee (2014). "Comparison of the effect of intra-tendon applications of recombinant human platelet-derived growth factor-BB, platelet-rich plasma, steroids in a rat achilles tendon collagenase model." Journal of Orthopaedic Research **32**(1): 145-150.

Spanoudes, K., D. Gaspar, A. Pandit and D. Zeugolis (2014). "The biophysical, biochemical, and biological toolbox for tenogenic phenotype maintenance in vitro." Trends in Biotechnology **32**(9): 474-482.

Steinert, A., S. Ghivizzani, A. Rethwilm, R. Tuan, C. Evans and U. Nöth (2007). "Major biological obstacles for persistent cell-based regeneration of articular cartilage." Arthritis research & therapy **9**(3): 213.

Sugimoto, Y., A. Takimoto, H. Akiyama, R. Kist, G. Scherer, T. Nakamura, Y. Hiraki and C. Shukunami (2013). "Scx+/Sox9+ progenitors contribute to the establishment of the junction between cartilage and tendon/ligament." Development (Cambridge, England) **140**(11): 2280-2288.

Sun, J. and H. Tan (2013). "Alginate-Based Biomaterials for Regenerative Medicine Applications." Materials **6**(4): 1285-1309.

Szallasia, Z., J. Stelling and V. Periwal (2010). System modeling in cellular biology: From concepts to nuts and bolts, MIT Press.

Takimoto, A., M. Oro, Y. Hiraki and C. Shukunami (2012). "Direct conversion of tenocytes into chondrocytes by Sox9." Experimental Cell Research **318**(13): 1492-1507.

Tan, C., P. P. Y. Lui, Y. W. Lee and Y. M. Wong (2014). "Scx-transduced tendon-derived stem cells (tdscs) promoted better tendon repair compared to mock-transduced cells in a rat patellar tendon window injury model." PloS ONE **9**(5).

Tarca, A. L., S. Draghici, P. Khatri, S. Hassan, P. Mittal, J.-S. Kim, C. J. Kim, J. P. Kusanovic and R. Romero (2009). "A novel signaling pathway impact analysis." Bioinformatics (Oxford, England) **25**(1): 75-82.

Thomopoulos, S., R. Das, M. Silva, S. Sakiyama-Elbert, F. Harwood, E. Zampiakis, M. Kim, D. Amiel and R. Gelberman (2009). "Enhanced flexor tendon healing through controlled delivery of PDGF-BB." Journal of Orthopaedic Research **27**(9): 1209-1215.

Thorpe, C., H. Birch, P. Clegg and H. Screen (2013). "The role of the non-collagenous matrix in tendon function." International Journal of Experimental Pathology **94**(4): 248-259.

Thorpe, C., C. Udeze, H. Birch, P. Clegg and H. Screen (2012). "Specialization of tendon mechanical properties results from interfascicular differences." Journal of The Royal Society Interface **9**(76): rsif20120362-20123117.

Turing, A. M. (1952). "The Chemical Basis of Morphogenesis." Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences **237**(641): 37-72.

van der Rest, M. and R. Garrone (1991). "Collagen family of proteins." The FASEB Journal **5**(13): 2814-2823.

van Dijk, G., C. Veenhof, F. Schellevis, H. Hulsmans, J. Bakker, H. Arwert, J. Dekker, G. Lankhorst and J. Dekker (2008). "Comorbidity, limitations in activities and pain in patients with osteoarthritis of the hip or knee." BMC Musculoskeletal Disorders **9**(1): 95.

Vasiliadis, H. and J. Wasiak (2010). "Autologous chondrocyte implantation for full thickness articular cartilage defects of the knee." The Cochrane Database of Systematic Reviews **10**: CD003323.

Vaske, C., S. Benz, Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler and J. Stuart (2010). "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM." Bioinformatics (Oxford, England) **26**(12): i237-i245.

Voleti, P., M. Buckley and L. Soslowsky (2012). "Tendon healing: repair and regeneration." Annual Review of Biomedical Engineering **14**: 47-71.

Wang, J. and A. Levchenko (2009). "Microfluidics technology for systems biology research." Methods in Molecular Biology (Clifton, N.J.) **500**: 203-219.

Wasiak, J., C. Clar and E. Villanueva (2006). "Autologous cartilage implantation for full thickness articular cartilage defects of the knee." The Cochrane Database of Systematic Reviews **3**: CD003323.

Wasiak, J. and E. Villanueva (2002). "Autologous cartilage implantation for full thickness articular cartilage defects of the knee." The Cochrane Database of Systematic Reviews **4**: CD003323.

Worthley, Daniel L., M. Churchill, Jocelyn T. Compton, Y. Tailor, M. Rao, Y. Si, D. Levin, Matthew G. Schwartz, A. Uygur, Y. Hayakawa, S. Gross, Bernhard W. Renz, W. Setlik, Ashley N. Martinez, X. Chen, S. Nizami, Heon G. Lee, H. P. Kang, J.-M. Caldwell, S. Asfaha, C. B. Westphalen, T. Graham, G. Jin, K. Nagar, H. Wang, Mazen A. Kheirbek, A. Kolhe, J. Carpenter, M. Glaire, A. Nair, S. Renders, N. Manieri, S. Muthupalani, James G. Fox, M. Reichert, Andrew S. Giraud, Robert F. Schwabe, J.-P. Pradere, K. Walton, A. Prakash, D. Gumucio, Anil K. Rustgi, Thaddeus S. Stappenbeck, Richard A. Friedman, Michael D. Gershon, P. Sims, T. Grikscheit, Francis Y. Lee, G. Karsenty, S. Mukherjee and Timothy C. Wang (2015). "Gremlin 1 Identifies a Skeletal Stem Cell with Bone, Cartilage, and Reticular Stromal Potential." Cell **160**(1–2): 269-284.

Yoon, B., D. Ovchinnikov, I. Yoshii, Y. Mishina, R. Behringer and K. Lyons (2005). "Bmpr1a and Bmpr1b have overlapping functions and are essential for chondrogenesis in vivo." Proceedings of the National Academy of Sciences of the United States of America **102**(14): 5062-5067.

Young, A., M. Smith, S. Smith, M. Cake, P. Ghosh, R. Read, J. Melrose, D. Sonnabend, P. Roughley and C. Little (2005). "Regional assessment of articular cartilage gene

expression and small proteoglycan metabolism in an animal model of osteoarthritis." Arthritis Research & Therapy **7**(4): R852-861.

Young, M. (2012). "Stem cell applications in tendon disorders: a clinical perspective." Stem cells international **2012**.

Zakany, J. and D. Duboule (2007). "The role of Hox genes during vertebrate limb development." Pattern Formation and Developmental Mechanisms **17**(4): 359-366.

Zeller, R., J. López-Ríos and A. Zuniga (2009). "Vertebrate limb bud development: moving towards integrative analysis of organogenesis." Nature Reviews. Genetics **10**(12): 845-858.

2: Convergent transcriptomic profiles arise from the dedifferentiation of cartilage and tendon cells in monolayer

Abstract

Dedifferentiation, a loss of cellular functionality, is a term with no definitive mechanism yet is used to described distinct and diverse biological contexts ranging from the histological features of neoplasia to regenerative responses in injury models. Recent evidence also implicates the loss of differentiated status as a factor in degenerative and chronic disease. In chondrocyte biology dedifferentiation is a well-recognised sequelae to expansion in monolayer culture. Regenerative therapeutics, such as autologous cartilage implantation, by necessity require periods of monolayer culture expansion. Consequently, exploration of mechanisms, which may contribute to degenerative phenotypes, could help elucidate points for therapeutic intervention in conditions such as osteoarthritis.

Despite considerable efforts to characterise soluble factors and environmental and mechanical cues that may actively promote differentiation of permissive cells towards a chondrogenic or tenogenic lineage the benchmarks by which progress is measured are inconsistent and limited. To date there is no available reference dataset against which progress in organo-typic culture techniques may be compared at a global transcriptomic level.

In this study there gene expression profiles native, monolayer and standard three-dimensional culture systems using chondrocytes and tenocytes are compared using

microarrays providing the first comparative data set of its kind. Results demonstrate an inadequate restitution of native tissue expression profiles by commonly used three-dimensional culture models. In addition, convergence of gene expression profiles in monolayer culture, and the expression of development-associated genes in these cells, implicates dedifferentiation as a mechanism worthy of further investigation or de- and re-generation of musculoskeletal tissues. In particular, the expression of a hind-limb development-associated homeobox gene, *Pitx1*, in monolayer chondrocytes is validated that suggests further investigation of homeobox genes in dedifferentiation is warranted.

2.1: Introduction

2.1.1: Plasticity of terminally differentiated cells

Convention presents us with the linear narrative of development and differentiation of cells from the pluripotent state to the terminal, functional, differentiated state in adult tissue. This trajectory follows binary cell-fate decisions implicit in the visual metaphor of Waddington's epigenetic landscape (Zhou and Huang 2011, Ferrell 2012). It is clear, through the use of somatic cell nuclear transfer (Campbell, McWhir et al. 1996) and induced-pluripotent stem cells (Takahashi and Yamanaka 2006), that the differentiated state is not an irreversible one and adult somatic cells may be phenotypically plastic in certain contexts. It is suggested, however, that the more specialist the cellular function the more difficult it is to reverse that differentiated state (Holmberg and Perlmann 2012). This is complemented by the concept of cells residing within an epigenetic landscape consisting of kinetic barriers and 'sinks' of attraction that ensure differentiated

states prevail in the adult (Enver, Pera et al. 2009, Huang 2009, Bhattacharya, Zhang et al. 2011).

Contemporary studies utilise terms such as de-, trans-, and re-differentiation to represent alterations in the synthetic profile of adult somatic cells in response to injury, ageing or disease (Poss 2010, Jopling, Boue et al. 2011). The mechanism of 'dedifferentiation', an organised loss of differentiated function, has been investigated in multiple novel model species from *Hydra* to the zebrafish (*Danio rerio*) (Grafi 2004, Sugimoto, Gordon et al. 2011). In addition to plant, amphibian, and fish models, mammalian models have been recently presented (Lehoczky, Robert et al. 2011, Nagoshi, Shibata et al. 2011, Porrello, Mahmoud et al. 2011). Dedifferentiation has also been considered as a target mechanism to separate the biological (functional) and chronological age of a cell (Rando and Chang 2012). However, the concept of dedifferentiation as an exclusive regenerative mechanism is conflicted by evidence supporting the presence of specialised populations of adult tissue stem cells as the principle effectors of regeneration (Sugimoto, Gordon et al. 2011). To clarify, the presence, or absence, of dedifferentiation as a regenerative mechanism in vertebrates relies on how dedifferentiated a cell is considered to be, i.e. the degree of 'stemness'. The extent of phenotypic plasticity in higher vertebrates is suggested to be much more restricted than would be expected of an embryonic stem cell. As such dedifferentiation is an ambiguous term, without mechanistic definition, and is used permissively in several contexts.

2.1.2: Definition of dedifferentiation varies with context

De-differentiation in cartilage biology

To cartilage matrix biologists the concept of cartilage dedifferentiation is well recognised (Schulze-Tanzil 2009) and is a term used with impunity for several decades. The morphological alteration from rounded chondrocytes in cartilage to a fibroblastoid phenotype is an inevitable response to two-dimensional monolayer culture (Von der Mark, Gauss et al. 1977, Benya and Shaffer 1982). Under these conditions chondrocytes rapidly lose their rounded morphology and functional phenotype, characterised by the pronounced down-regulation of collagen type II, and aggrecan, hallmarks of cartilage tissue, with passage four considered a threshold to terminal dedifferentiation (Darling and Athanasiou 2005, Schulze-Tanzil 2009). In their seminal study Benya and Shaffer (1982) explicitly state that 'dedifferentiated', relating to chondrocytes in monolayer, referred only to a loss of differentiated function and not evidence of a multipotent cell (Benya and Shaffer 1982), although no evidence was provided to support this assertion.

De-differentiation as a regenerative mechanism

Dedifferentiation is also used to describe a response in adult cells to injury as demonstrated by regeneration models such as axolotl forelimb amputation. In these models it is implicit that adult cells lose their functional gene expression profile, up-regulate expression of genes associated with an earlier developmental stage, re-enter the cell cycle and proliferate, before recapitulating development in a tissue lineage and position-specific manner (Kragl, Knapp et al. 2009). Recently contrasting results indicated that dedifferentiated myofibres proximal to a forelimb amputation site made no contribution to the proliferating cell population in the regenerating limb of the axolotl, but made a significant contribution to

regeneration in the newt, in a cell tracing study (Sandoval-Guzmán, Wang et al. 2014). Further to this, a study by Tata, *et al* (2013) found that luminal epithelial secretory cells of mouse airways dedifferentiated to, and were indistinguishable from, basal stem cells, following conditional ablation of the resident progenitor population (Tata, Mou et al. 2013).

De-differentiation as a pathological mechanism in chronic disease

Insulin-secreting pancreatic beta (β) cells also dedifferentiate when expanded in culture and have been shown to return to a functional, insulin-secreting state in certain culture conditions (Russ, Sintov et al. 2011). Recently dedifferentiation has been demonstrated *in vivo* as a mechanism for beta-cell failure in a model of Type II diabetes (Talchai, Xuan et al. 2012). Evidence suggests that a transient period of dedifferentiation is necessary prior to proliferation in response to β-cell loss (El-Gohary, Tulachan et al. 2014).

Dedifferentiation has been described in cardiomyocytes where it is characterised by a loss of electrophysiological properties, disassembly of the sarcomeric structure, proliferation, and expression of progenitor gene markers. Cardiomyocyte dedifferentiation is now well described *in vitro* (Zhang, Li et al. 2010) and *in vivo* (Dispersyn, Mesotten et al. 2002, Jopling, Sleep et al. 2010, Porrello, Mahmoud et al. 2011) and more recently as a key feature of cardiomyocyte remodeling (Szibor, Pöling et al. 2014). Dedifferentiation also has a characterised association with chronic cardiac pathologies (Ausma, Wijffels et al. 1997, Driesen, Verheyen et al. 2009).

Common to these regenerative models are i) loss of functional markers, ii) periods of proliferation, iii) evidence of tissue-restricted progenitors, and iv) and function-

restoring redifferentiation processes. It is not illogical to suggest that cells that have dedifferentiated, no longer expressing the synthetic profile of the specialist cell, would fail to contribute to tissue function. Whether dedifferentiation contributes to chronic, degenerative disease in musculoskeletal tissues, or a regenerative response, has yet to be elucidated.

Dedifferentiation in neoplasia

Dedifferentiation in neoplasia is used to describe the histopathological changes associated with deregulated tumour masses. It has been demonstrated that NF-κB modulation of Wnt-signalling can lead to the dedifferentiation of intestinal epithelial cells not of stem cell origin into cells with a crypt-progenitor phenotype, and subsequently into tumour-initiating cells (Schwitalla, Fingerle et al. 2013). Dedifferentiation of epithelial cells towards a tumourigenic status has also been described from neurons during glioma formation (Friedmann-Morvinski, Bushong et al. 2012), but is also used to describe the development of 'cancer stem cells' as a self-renewing source distinct from cancer cells derived through a Wnt-/β-catenin pathway (Debeb, Lacerda et al. 2012). Signalling pathways associated with poorly differentiated cells in cancers may be relevant to an understanding of the underlying mechanisms involved in dedifferentiation.

Summary

Dedifferentiation of somatic cells in physiological conditions represents some loss of the functional characteristics of the differentiated cell. In the adult mammal this is unlikely to represent a cell with multi-lineage potential as found in the amphibian. As a term dedifferentiation is applied to regeneration, degeneration, and neoplasia, but it is not clear whether the same regulatory mechanism is in

place in these contexts. Although dedifferentiation is recognised in cartilage biology there has been no systematic investigation of this process.

2.1.3: Anatomical topography in development is encoded by homeobox genes

Homeobox (*Hox*) genes contain homeobox DNA sequences and encode transcription factors that specify positional identity during embryonic development (Wang, Helms et al. 2009); the proteins contain the 'homeodomain'. Positional identity has been likened to a Cartesian coordinate system or 'genome GPS' for cells, defined by the expression signatures of the ~300 homeobox genes (Chang 2009). *Hox* gene expression codes have been found to define fibroblast (Rinn, Bondre et al. 2006) and mesenchymal stem cell populations (Ackema and Charite 2008, Sagi, Maraghechi et al. 2012) based on the tissue and site of origin.

Loss of cellular identity through alterations in *Hox* gene expression (Trivedi, Cappola et al. 2011) or cellular dedifferentiation to a progenitor-like state appears to result in aberrant cellular homeostasis and age-related functional deterioration. *Hox* gene-mediated transcriptional memory appears to limit stem cell-mediated tissue regeneration (Chang 2009). Whilst *Hox* gene function has been well studied during embryonic segmentation, there is limited information on gene expression patterns in adult tissues.

Summary

Given the relevance of *Hox* gene expression to differentiated status, and positioning during development, altered expression patterns may be evident between differentiated and dedifferentiated cells.

2.1.4: Study aims

This study hypothesises that the gene expression profiles in de- and re-differentiation of chondrocytes and tenocytes in monolayer and three-dimensional cultures are consistent with many of the criteria associated with dedifferentiation in the context of regeneration and disease models.

In addition to a return to proliferation, which is a prerequisite of monolayer expansion, there is a) a loss of a broad range of differentiation markers, b) re-expression of genes specifically associated with tissue development, c) evidence of positional identity, c) restricted potential, and d) recapitulation of development when dedifferentiated cells are placed in a supportive, three-dimensional environment.

The study sought to develop the profile of known markers of differentiated status in cartilage and tendon; define a dedifferentiation profile; establish features of dedifferentiation that may be consistent with established markers of degeneration of cartilage and tendon; and compare the expression profile of three-dimensional culture systems relative to the native tissue they attempt to model. Additionally, the expression profile of *Hox* genes in adult cartilage and tendon are defined to establish whether musculoskeletal tissue in the adult retains expression of markers of positional identity.

This study sevidence of global transcriptomic changes associated with dedifferentiation, including expression of lineage-associated development markers and alteration in the expression of *Hox* genes, but also a failure of three-dimensional culture systems to faithfully restore differentiated status. These findings provide supportive evidence for the plasticity of dedifferentiated

musculoskeletal cells and the need for further exploration of dedifferentiation in organotypic culture systems.

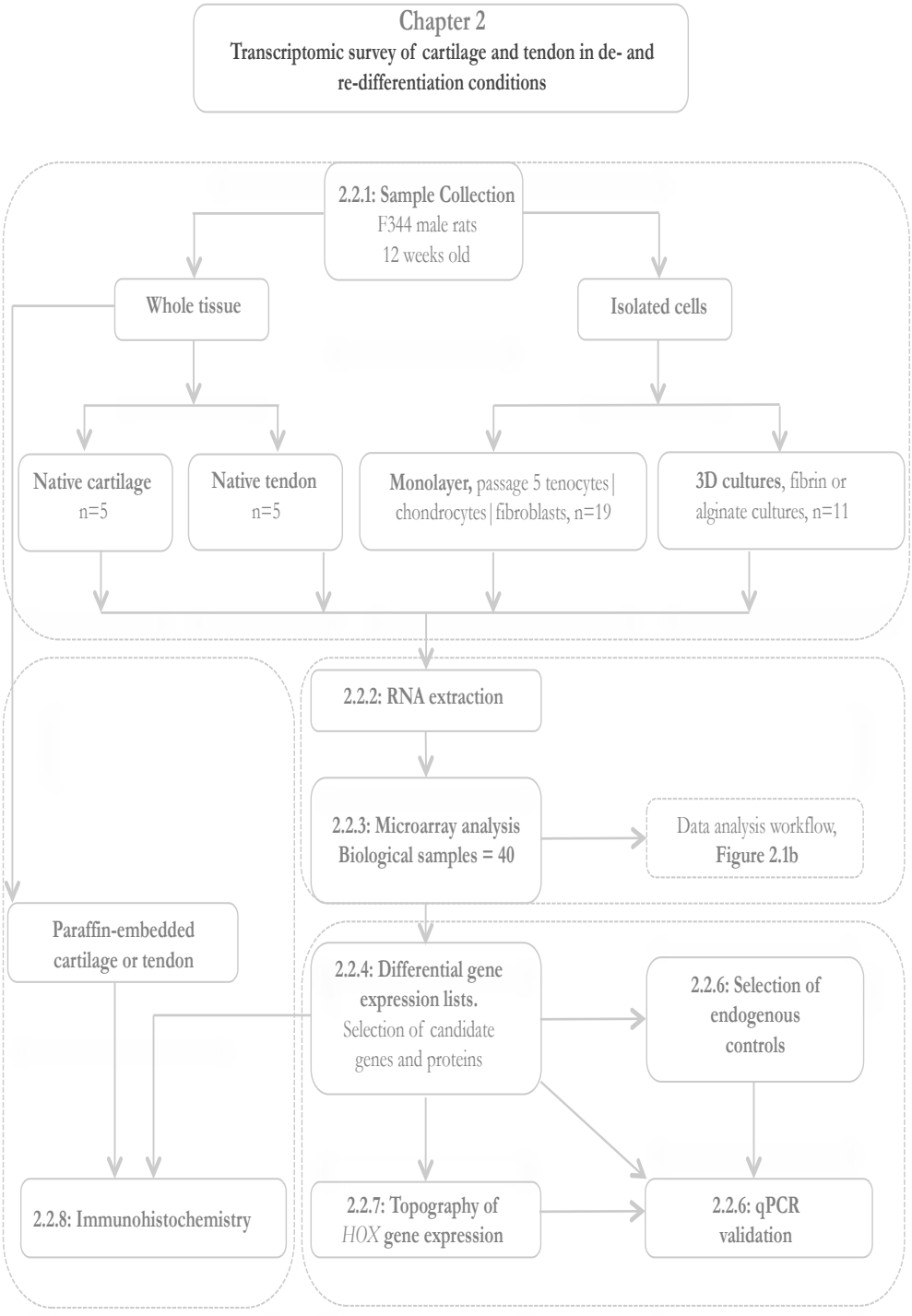


Figure 2.1a: Overview of experimental design for results presented in **Chapter 2**. Forty microarrays were used to interrogate samples derived from n=18 biological replicates. Two tissue sources and three environmental conditions are shown – native, monolayer and three-dimensional cultures. Relevant sections in the methods are provided.

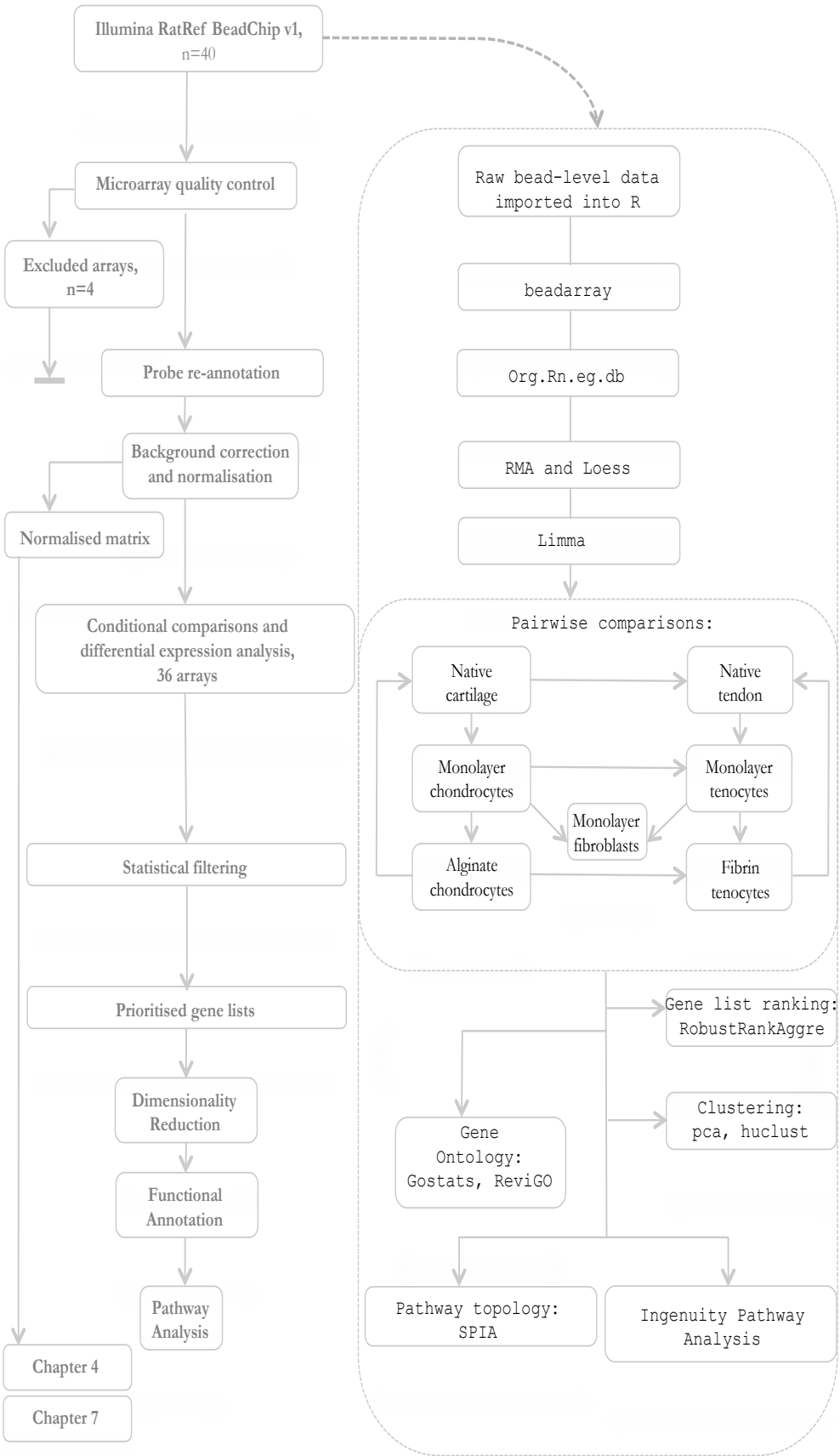


Figure 2.1b: Data analysis workflow (left column) and bioinformatics pipeline (right column) for data arising from Illumina microarrays. Pairwise comparisons between conditions are shown within the pipeline with the direction of the comparison indicated by the arrow. *R* packages or functions used are also provided.

2.2: Methods

A graphical overview of the workflow employed in this chapter and sample sources is provided in **Figure 2.1a** and **2.1b**.

2.2.1: Culture protocols

Tissue

Tissue was obtained from twelve week old, skeletally mature (Roach, Mehta et al. 2003) male F344 rats (F344/IcoCR (Charles River, n=5) or F344/NCrHsd (Harlan Laboratories, Inc, n=13)), total n=18 (mean weight±s.d; 248 g ± 25.5). All rats were communally housed for seven days after transport and were humanely destroyed in compliance with the Animals (Scientific Procedures) Act 1986, Schedule I. Tissue was harvested following a short delay *post-mortem* (3.5 hours ± 1.4) (Marchuk, Sciore et al. 1998). All cartilage and tendon tissue was harvested from the hind limbs. Cartilage was pooled from coxo-femoral (hip) and femorotibial (knee) joints; tendon was pooled from the tendon of the gastrocnemius (Achilles), tail and the deep flexor tendons of the hind limbs. From the left flank an area of approximately 2 cm^2 was obtained for isolation of dermal fibroblasts. Cartilage was digested in 0.2 % collagenase type II (Worthington, Lakewood, NJ, USA) at 37 °C for 12-18 hours in 15 mL conical Falcon tubes (Falcon, BD Biosciences). All cell culture reagents were from Gibco, (Life Technologies, Carlsband, CA, USA) unless otherwise stated. Tendon and dermis were pre-digested with 0.25 % (w/v) trypsin for 20 minutes, with further digestion as described for cartilage. For primary culture, cells were seeded at 10^3 cells/cm^2 in 25 cm^2 cell-culture flasks with low glucose Dulbecco's Modified Eagle's Medium (DMEM) with L-glutamine, with the addition of 10 % foetal calf serum (Sigma-Aldrich, St. Louis, MO, USA), 1 % penicillin (100 U/mL) and

streptomycin (100 µg/mL), and 0.2 % amphotericin B (2 µg/mL), which will be referred to as culture medium 1 (CM1). Cells were incubated at 37 °C in a humidified environment (5 % CO_2: 21 % O_2). Growth medium was changed every 2-3 days to maintain active proliferation. Cells were grown to 80-90 % confluence. Prior to subculture the cell monolayer was washed with 10 mL of HBSS. For subculture cells were dissociated from flasks using a 0.05 % trypsin-EDTA solution (1 mL per 75 cm^2), followed by a cell count using a modified Fuchs-Rosenthal method and trypan-blue (0.4 %, Sigma-Aldrich) exclusion test on 10 µL samples. Over 90% of cells were viable at all passages. The first passage occurred once the primary cell culture reached confluence. For all subsequent subcultures cells were seeded at 10^4 cells/cm^2. Cells underwent subculture on four further occasions and were harvested at passage five. Population doublings for each culture did not reach the limit defined by Hayflick (Hayflick 1961).

Alginate beads cultures

Passage five chondrocytes, or tenocytes, were released from monolayer culture, pelleted at 500 x *g* for 10 minutes and washed twice in sterile phosphate-buffered saline (PBS) after discarding the supernatant. Cells were re-suspended at a density of $2x10^6$ cells/mL in 1.2 % sterile-filtered alginic acid (alginate) solution (Sigma-Aldrich, as before) adapted from de Ceunicnk, *et al* (2004) (De Ceuninck, Lesur et al. 2004). CM1, with the addition of filter-sterilised L-ascorbic acid-2-phosphate at a final concentration of 200 µM (Sigma-Aldrich), is referred to as culture media 2 (CM2). Beads were incubated in 100 mm diameter Petri dishes in 25 mL of CM2 for 14 days, as before, with media changed twice weekly. A cell viability study defined the number of trypan-blue positive (dead) cells taken from individual

alginate beads on six occasions over fourteen days (n=4, triplicate counts for each replicate) for both chondrocytes and tenocytes.

Fibrin constructs

Fibrin culture systems were prepared using a previously described technique (Kapacee, Richardson et al. 2008). The CM2 media for alginate beads was modified by the addition of aprotinin (20 μg/mL (Aprotinin from bovine lung, Sigma-Aldrich, A1153), after Ye, *et al* (2010), to inhibit fibrinolysis (Ye, Zünd et al. 2000). In each well, 7.5 x10^5 cells (using the modified cell number in (Kapacee, Yeung et al. 2010)) passage 5 tenocytes or chondrocytes were suspended in filter-sterilised 20 mg/mL fibrinogen and 200 U/mL thrombin (Sigma-Aldrich) with modified CM2 to a final volume of 480 μL; this was added to each well and was rapidly agitated to ensure even coverage of the well. Wells were left at 37 °C for five minutes to permit a thin fibrin layer to form before the addition of 5 mL of modified CM2. The fibrin layer was scored circumferentially with sterile 10 μL pipette tip after 24 hours, and then scored centripetally every other day, to facilitate contraction of the developing construct. Constructs were incubated, as before, until a robust linear construct was formed after 7-10 days.

2.2.2: RNA extraction

Monolayer cells were washed twice with HBSS then lysed using a choatropic agent (TRIzol®, Ambion, as before) and left for 10 minutes at room temperature. RNA isolation was then performed using an acid guanidinium thiocyanate–phenol–chloroform extraction method (Chomczynski and Sacchi 2006). Following co-precipitation with glycogen and re-suspension with 75 % ethanol (v/v) RNA was transferred to spin-columns to undergo an on-column DNAse digest, wash and purification (Qiagen GmbH, Hilden, Germany) as per the manufacturer's

instructions. Native tissue samples were minced in petri dishes and stored in ten volumes of RNAlater (Ambion, as before) and handled as per the manufacturer's instructions. For extraction, RNAlater was removed and the tissue snap-frozen in liquid nitrogen and pulverised using either a sterile mortar and pestle (tendon) or dismembrator (cartilage). Pulverised tissue was then incubated at room temperature in TRIzol® before undergoing the RNA extraction protocol described above. All samples were quality controlled and quantified using a spectrophotometer (Nanodrop, Thermo Scientific). RNA was eluted in DEPC-treated dH_20 and stored at -80 °C until required.

Cells were recovered from alginate beads by incubation with a depolymerising solution, to chelate calcium, (55 mM sodium citrate, 150 mM sodium chloride, pH 6) in a shaking incubator at 37 °C for ten minutes. The solution was centrifuged at 500 x *g* for ten minutes. The supernatant was discarded and the cell pellet was washed twice in warmed PBS and centrifuged as before. RNA extraction proceeded as described above on the isolated cell pellet. For fibrin constructs media was aspirated and constructs washed twice *in situ* with warmed PBS. Constructs were sharply dissected from the two anchor points using a scalpel blade and transferred directly to a chaotropic agent.

2.2.3: Microarray analysis

Microarray samples were derived from several sources: a) native (whole tissue) cartilage or tendon, b) monolayer-expanded chondrocytes, tenocytes or dermal fibroblasts at passage five, or c) cells derived from alginate or fibrin three-dimensional cultures. Forty microarrays were analysed, **Table 2.1**, using the Illumina RatRef-12 v1.0 BeadChip® array (Illumina, Inc., San Diego, California, USA) (Oliphant, Barker et al. 2002) following submission to The Genome Centre,

Barts and the London, Queen Mary, University of London. Twelve samples could be run in parallel on each array and were performed in three batches (native and monolayer in two batches, with three-dimensional cultures as a separate batch) on separate occasions. On each array there were 22,523 randomly distributed gene-specific bead probes for the rat reference genome. RNA quality and quantity was assessed prior to amplification using the Agilent Bioanalyser RIN system (Agilent Technologies). Samples with a RIN-value greater than 8 were considered to have passed quality control. Labelling was undertaken using standard manufacturers protocols for RNA amplification based upon the technique by Van Gelder, *et al* (1990) (Van Gelder, von Zastrow et al. 1990). Biotinylated cRNA was prepared with Illumina TotalPrep® RNA amplification kit (Ambion). Hybridisation and image acquisition were performed using the manufacturer's standard protocol for this array. Raw data text files and un-normalised expression data were obtained from Illumina BeadStudio® software output and used at the input for data analysis, **Figure 2.1b**.

Condition	Replicates	Comments
Cartilage	5(4)	Pooled hip and knee cartilage, n=1 removed at quality control (QC)
Tendon	5(5)	Pooled Achilles, tail and deep flexor tendon
Chondrocytes (monolayer	8(8)	Passage 5
Tenocytes (monolayer)	8(8)	Passage 5
Fibroblasts (monolayer	3(3)	Passage 5
Alginate (chondrocytes)	4(4)	-
Alginate (tenocytes)	1(0)	Removed at QC
Fibrin (tenocytes)	3(2)	One array removed at QC
Fibrin (chondrocytes)	3(2)	Chondrocytes transferred to fibrin constructs. One array removed at QC.

Table 2.1: Table demonstrates source of samples, biological replicates and the number of arrays actually used for analysis in parentheses, total n=**36**. For three-dimensional cultures the component cell type is also given.

2.2.4: Data Analysis

Quality control of microarray analysis

General analysis of raw and normalised data was performed to assess systematic and individual array errors that may have arisen during the scanning protocol. All analysis was performed using the *R* programming platform, *R version 3.0.2* (2013, The R Foundation for Statistical Computing) through open-source packages made available via the Bioconductor network, (http://www.bioconductor.org (Gentleman, Carey et al. 2004)). Data was loess-normalised (Bolstad, Irizarry et al. 2003) and $\log_2$-transformed after alternative techniques were investigated using the `beadarray` package, v2.6.0, (Dunning, Smith et al. 2007), implementing the

BASH algorithm to analyse spatial artifacts (Cairns, Dunning et al. 2008). This data analysis included the removal of four arrays at quality control that were considered unsuitable for further analysis (remaining for analysis, n=36 arrays).

Differential expression analysis

Statistical analysis of differential gene expression was performed using the `limma` package (Smyth 2004). Results for gene expression are presented as the $\log_2$ fold change ($\log_2$FC), false discovery rate (FDR, Benjamini-Hochberg correction) and log-odds ratio of expression (B statistic). Pairwise comparisons are described such that the first term defines the baseline condition to which a comparison is made, for example, 'native cartilage to monolayer chondrocytes' defines native cartilage as the baseline. Complete differential expression lists for pairwise comparisons are available in supplementary data **SD2.1-2.10**. Only differentially expressed genes passing a filtering threshold were used for bioinformatic analysis: $\log_2$ fold-change ($\log_2$FC) $\geq$0.5 (absolute fold-change =1.4); FDR <0.01; B statistic >0 (equivalent to a 50% likelihood of differential expression). Illumina identifiers were re-annotated using `org.Rn.eg.db`. (Pages H, Carlson M, Falcon S and Li N. AnnotationDbi: Annotation Database Interface. R package version 1.28.1).

2.2.5: Bioinformatics

All web-interface bioinformatics tools were last accessed in November 2014.

Dimensionality reduction

Principal component analysis (PCA) was undertaken using a co-variance matrix (Husson 2010) from filtered expression data from 36 arrays based upon the top 500 most co-variant genes establish using `WGCNA` (Langfelder and Horvath 2008).

Hierarchical clustering using the was based upon the complete linkage method (Murtagh 1985) using the same gene expression matrix.

Gene ontology function annotations

Gene ontology (GO) (Ashburner, Ball et al. 2000) functional annotation enrichment was assessed using a strict hypergeometric analysis with the package `GOstats` (Falcon and Gentleman 2007). Gene ontology analysis was also undertaken using DAVID (http://david.abcc.ncifcrf.gov) (Huang, Sherman et al. 2008), with appropriate species and platform backgrounds, for validation. Output GO lists were rationalised using ReviGO (http://revigo.irb.hr) (Supek, Bošnjak et al. 2011). The 'SimRel' algorithm was used to calculate the semantic similarity score. The UniProt *Rattus norvegicus* database (2013) was used to define the search space. Only terms with a FDR<0.001 were used.

Pathway Topology:

Canonical signalling pathways were obtained from The Kyoto Encyclopedia of Genes and Genomes, KEGG (http://www.genome.jp/kegg/), for *Rattus norvegicus* (58 pathways used) in XML format (Kanehisa and Goto 2000). Total, filtered differential expression lists consisting of Entrez gene identifiers and $\log_2$ fold-changes were used as the input to the `SPIA` pathway topology package in *R,* version 2.14.0 (Tarca, Draghici et al. 2009). For a pairwise comparison, for example native cartilage versus monolayer chondrocytes, positive $\log_2$ fold-changes represented higher expression in native cartilage.

Prioritised gene lists using rank aggregation

These gene lists consisted of two sets of effect sizes, $\log_2$FC, arising from either cartilage or tendon. In order to integrate differential expression results from

different comparisons into a meaningful prioritised gene list a rank aggregation method (Kolde, Laur et al. 2012) was employed. Ranked gene lists for each pair-wise comparison were ordered by ascending adjusted p-value (FDR).

Inference of Upstream Regulators of Gene Expression

To infer upstream master regulators of gene expression the Ingenuity® Pathway Analysis (IPA®, Qiagen Bioinformatics, Redwood City, USA, www.ingenuity.com) knowledge base and software implementing causal analysis methods (Krämer, Green et al. 2014) were used under license. Briefly, regulators with network connections to, and the direction of regulation within, the expression dataset were scored on their likelihood of occurring more frequently than in a random model. The top upstream regulators (including: transcription factors, small molecules, endogenous chemicals, miRNAs) were defined in this study as those with: i) the smallest 'overlap *p*-value' – a measure of enrichment of regulated genes within a dataset using a Fisher's Exact Test (right-tailed), and ii) the highest 'activation *z*-score' – the activation state of a regulator inferred from a test of the match in up- and down-regulation patterns. To build on the mechanistic networks predicted through IPA, downstream targets of transcription factors were collected from the existing gene expression dataset. Only those genes that were differentially expressed and had a direct relationship with the master regulators were chosen.

2.2.6: Validation techniques

Reverse transcription production of cDNA

Random hexamers were annealed to 1 µg RNA with addition of 0.2 mM of each dNTP in a 25 µL reaction, with M-MLV reverse transcriptase and buffer, and

RNAse inhibitor in volumes prescribed by the manufacturers (all Promega, Madison, WI, USA).

Determination of endogenous control genes for delta-C_t method of normalisation

A series of endogenous control genes were selected from a review publication (de Jonge, Fehrmann et al. 2007) and commercially available multiplex qPCR endogenous control arrays. These were cross-referenced with the mean $\log_2$ fluorescence intensity from all Illumina arrays. Genes with a coefficient of variation (CV) < 1% and a mean fold change < 2 across all tissues and culture conditions were chosen for the next round of validation. Quantitative PCR was undertaken using three biological replicates with technical triplicates for each. Raw C_t values were exported into the *R* package `NormqPCR` (Perkins, Dawes et al. 2012) which implements the geNorm algorithm (Vandesompele, De Preter et al. 2002) to define the least variable gene pairs. The ribosomal protein *Rps20* was used for all normalisation; there was no evidence of differential expression of this gene in any pairwise comparisons.

Quantitative polymerase chain reaction (qPCR)

Primers were designed for qPCR using the NCBI PrimerBlast tool (Ye, Coulouris et al. 2012) (http://www.ncbi.nlm.nih.gov/tools/primer-blast) against the most recent mRNA records and spanning exon-exon boundaries where possible. Annealing temperature of all primer pairs was 60 °C ± 0.2 °C. All primers had efficiencies > 98% based upon six serial ten-fold dilutions. Quantitative PCR was performed using a complete mix containing SYBR intercalating dye, ROX passive reference, UNG and dNTPs (MESA BLUE, Eurogentec, Germany) at the concentrations recommended by the manufacturer.

The standard protocol for all qPCR reactions performed using the 7300 ABI platform (Roche, Switzerland) consisted of: 50 °C (2 mins), 95 °C (10 mins) followed by 40 cycles of 95 °C (15 s) and 60 °C (1 min).

Specificity of PCR products on validated primers was determined by melt-curve analysis. The predicted molecular weight of PCR products was verified by electrophoresis relative to a molecular weight marker on a 1% agarose gel impregnated with ethidium bromide and visualised under UV light. All details for primers are presented in **SD2.11**. 'No template' and 'no reverse transcriptase' negative controls were also run in parallel for each sample.

Validation of genes differentially expressed in microarray studies

The baseline cycle threshold (C_t) for each qPCR run on 96-well plates was automatically generated using ABI software to ensure C_t threshold was within the linear phase of the exponential curve. cDNA was diluted 1:3 and 3 µL of cDNA was used for each reaction well to minimise pipetting errors. The comparative delta C_t method, as described by Schmittgen and Livak (Livak and Schmittgen 2001, Schmittgen and Livak 2008), was used. Technical triplicates for each of three biological replicates were averaged and were normalised to the means of technical replicates for *Rps20* for the same biological replicate.

Statistical analysis of qPCR data

Normalised and linear transformed qPCR data ($2^{\wedge -dCt}$) was tested for deviation from a Guassian distribution using the Shapiro-Wilk test within *R*. Analysis of variance (ANOVA) of group means was undertaken for qPCR validation data; *post hoc* Tukey multiple-comparison of means was performed with a 95% family-wise confidence level.

2.2.7: Expression of homeobox genes in adult cartilage and tendon

Six, twelve week old, male outbred (Lewis) rats (Harlan, as before) were harvested for cartilage and tendon from discrete anatomical sites in both the fore- and hind-limbs. Specifically these were: scapulo-humeral joint (shoulder), fore- and hind-digital flexor tendons, and the coxofemoral joint (hip). Samples were prepared as described for native tissue. Homeobox genes with established topographical and/or tissue-specific developmental expression in the literature were selected. A panel of 19 homeobox genes were considered. For each independent biological replicate qPCR reactions (technical triplicates) were performed over four separate 96-well plates with a standardised plate design. The qPCR protocol used was consistent with the description above. There were 2,304 data points in total. In order to deal with 'non-detects', or missing data cells, the following strategy was used to impute values from the available data cells for technical triplicates: a) where one technical triplicate was missing the mean of the remaining two was substituted; b) where the mean value of the remaining technical replicates was ≥ 35 (considered the limits of real-time qPCR sensitivity), or c) where two or more technical replicates were missing, the missing data point was considered to represent no amplified product and $C_t = 40$ was the imputed value. Across plates the median absolute deviation (MAD), used in the `beadarray` package, was employed as a measure of central tendency and was calculated for the summary values from each biological replicate (n = 6). A threshold of ± 2.5*MAD defined outlier values to be removed (Leys, Ley et al. 2013). As before, *Rps20* was used as the endogenous control and means of technical triplicates were normalised to sample-matched *Rps20* controls. For each gene there were six summary observations ($2^{\wedge -dCt}$) for each of four tissues. To account for the technical variation across 24 separate qPCR studies data was $\log_2$ transformed and

studentised such that the mean was equal to zero and the standard deviation equal to ± 1. This generated a data matrix comprising 'z-scores' and was further analysed in *R*.

Departures from a Guassian distribution were assessed using the Shapiro-Wilks test. Where outlier values were removed an alternative implementation of the ANOVA, using weighted means and type I sum of squares, was employed for unequal group sizes and *post-hoc* testing performed using the Games-Howell test, appropriate for unequal group sizes and heteroscedastic data.

2.2.8: Histology and Immunohistochemistry

Tendon tissue was harvested from 12 week old, male Lewis rats (**Chapter 3**) from the tendon of the gastrocnemius muscle (Achilles) and the deep flexor tendon. Achilles tendons were divided into proximal and distal sections, whilst the deep flexor tendons were sectioned in three parts (origin, mid-portion, and insertion). Where possible a small portion of skeletal muscle tissue was retained to act as an internal positive control. Tissue samples were placed in fresh 4% para-formaldehyde and stored at 4 °C for 7 days. Tendon samples were paraffin-embedded in a longitudinal orientation and 5 micron sections obtained.

For analysis of cartilage, whole knee samples, consisting of the distal one third of the femur, femorotibial joint, and proximal one third of the tibia, were fixed as described and decalcified in 12.5% EDTA/1.25% NaOH (w/v, pH 7.4) for six weeks. Embedding, sectioning and heamatoxylin and eosin (H&E) or Masson's Trichrome staining was undertaken as a service by the Pathology Department of the School of Veterinary Science, University of Liverpool. Special acknowledgement is extended to Ms. Valerie Tilston for this work.

For immunohistochemistry paraffin wax was removed through sequential ten-minute baths in xylene and graded ethanol stages, followed by rehydration in deionized water. All reagents were supplied by Sigma-Aldrich (catalogue numbers provided), as before, unless otherwise stated. Antigen retrieval for both cartilage and tendon sections was undertaken with chondroitinase ABC (5 U/g) (#C2905) with 5 μL/mL in a Tris (50 mM)|sodium acetate (60 mM) buffer (pH 8 at romm temperature (RT)) according to the manufacturer's guidelines and (Linhardt 2001). Each slide was incubated with 200 μL of the enzymatic solution for twenty minutes at 37 °C. Subsequently, endogenous peroxidases were quenched in a 3% peroxide solution (v/v) for ten minutes, except for 3,3'-diaminobenzidine (DAB)-only controls, and then washed in de-ionised water. Alternative techniques employing pre-digestion with a topical solution of 0.05% trypsin from bovine pancreas (w/v, #T1426) in $CaCl_2$ at pH 7.8 for 20 minutes were also attempted.

Samples were washed in 1xTBS (pH 7.6 at 25 °C) with 0.1% TWEEN20 (v/v, TBST buffer). Sections were delimited using a hydrophobic pen, and blocked using 5% normal serum (v/v, diluted in TBST) specific to the species in which the secondary antibody was raised, either goat (#G9023) or donkey (#D9663), for one hour at 25 °C.

Blocking buffer was removed and replaced with 100-200 μL of primary antibody diluted in the blocking buffer, **Table 2.2**, and incubated overnight (> 12 hours) at 4 °C. Sections were washed three further times in TBST before the application of either: i) an anti-sheep HRP-polymer conjugated secondary antibody (GBI Labs Inc., Bothwell, USA, #D85-6) incubated for ten minutes following the application of an enhancer; or, ii) an anti-rabbit HRP-polymer (Zytomed Systems GmbH,

Berlin, Germany, #ZUC032-006) and incubated in a humidified atmosphere at 25 °C for a further sixty minutes.

Samples were washed twice in ultrapure dH_20 for five minutes before the application of 3,3'-diaminobenzidine (DAB) (SIGMAFAST™ tablets, #D4293) and colour allowed to develop over five minutes. Sections were immersed in dH_20 for five minutes to retard any further colour development. Sections were counter-stained with hematoxylin solution according to Delafield (#03971) for ten seconds before being left under running tap water for five minutes. Sections were dehydrated through graded ethanol and xylene, air-dried and mounted using DPX mountant (#06522). Quadriceps muscle or myotendinous junction served as a positive tissue control for TNNI2; bone marrow canals and blood vessels in knee sections served as positive tissue controls for CRAMP. Species-specific IgG isotype controls were concentration matched to test primary antibodies. No primary antibody (secondary only) and a DAB only, for endogenous peroxidase activity, were also included.

Target	Antibody Description	Source	Stock (Dilution)
TNNI2, troponin	Sheep, anti-rabbit polyclonal IgG	Abcam, UK (ab97711)	2 mg/mL (40 ug/mL)
CRAMP/CAMP	Rabbit, anti-mouse polyclonal IgG	Innovagen, Sweden	1 mg/mL (20 ug/mL)
Sheep IgG	Non-immune serum	Santa Cruz Biotechnology (sc-2717)	400 ug/mL (40 ug/mL)
Rabbit IgG	Non-immune serum	Santa Cruz Biotechnology (sc-2027)	400 ug/mL (20 ug/mL)
Anti-sheep	HRP-conjugated polymer (PoLink-2 Plus)	GBI Labs (D85-6)	As per manufacturers instructions
Anti-rabbit	HRP-conjugated polymer	Zytomed Systems	As per manufacturers instructions

Table 2.2: Antibody sources for immunohistochemistry studies

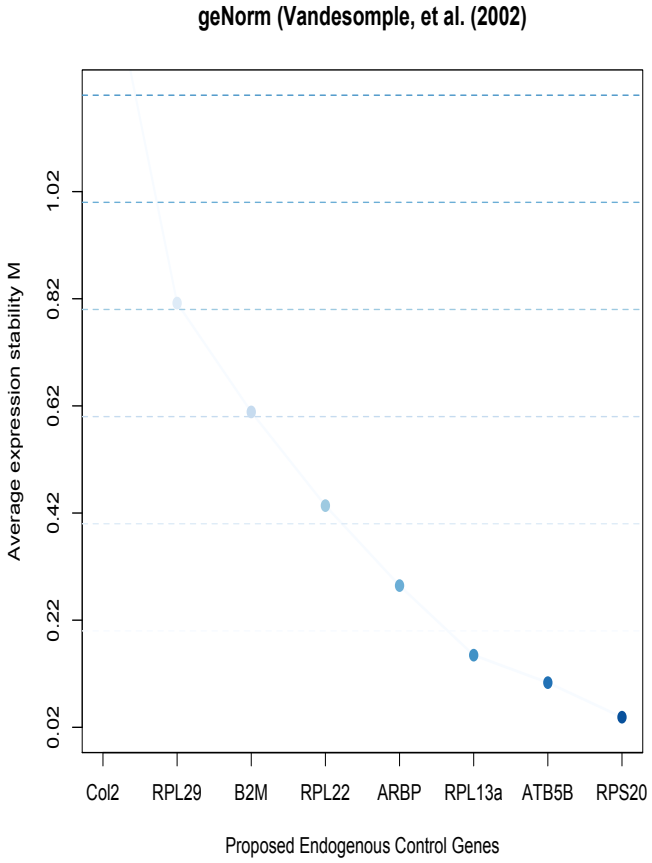


Figure 2.2: Plot of proposed endogenous control genes relative to the average expression stability value, M. Of the those tested *Atb5b* and *Rps20* were defined as the least variable across samples; it is not possible to rank the two most stable genes using this technique as it is based on the gene ratios. All normalisation was undertaken using *Rps20*.

2.3: Results

2.3.1: *Rps20* is a suitable endogenous control gene

After global assessment of gene expression variation for all microarray *Rps20* and *Atb5b* were found to be the most invariant, **Figure 2.2**. For normalisation of qPCR data across samples *Rps20* was used. Standard endogenous control genes, *Gapdh* and *Actb*, were differentially expressed between conditions and were not suitable candidates.

2.3.2: Chondrocyte and tenocyte gene expression profiles converge in monolayer culture

Principal Component Analysis (PCA)

After filtering on the top 500 most covariant genes derived from a subset of highly variant genes (> 0.8) arrays could be clustered into four distinct groups, **Figure 2.3**. The first two components described 70.2% of the variation in the data. These groups consisted of: i) native cartilage; ii) native tendon; iii) monolayer cultured chondrocytes, tenocytes, and fibroblasts; iv) or three-dimensional culture systems, (alginate or fibrin cultures). It was not possible to discriminate between cell types for monolayer or three-dimensional culture conditions. Native tendon and cartilage were strongly divergent from each other and cultured cells. Unsupervised PCA, without filtering of invariant genes, was poorly discriminatory (> 0.2, 43.9% described by the first two components); the overall relationships described were present with less stringent filtering, however, clustering of native samples was less robust. Inclusion of a third component did not discriminate further between the groups.

The variation in the gene expression profile was suggestive of a convergence of expression profiles for cells in monolayer and a failure of chondrocytes or tenocytes in three-dimensional culture to recapitulate differentiation status to parity with native tissue.

Hierarchical Clustering

Unsupervised hierarchical clustering of the top 500 most co-variant genes concurred with principal component analysis and delineated the data into clades of cartilage, tendon, 3D culture systems and monolayer-expanded cells, **Figure 2.4**. The genes that comprise the top 500 most co-variant are listed in **SD2.12**.

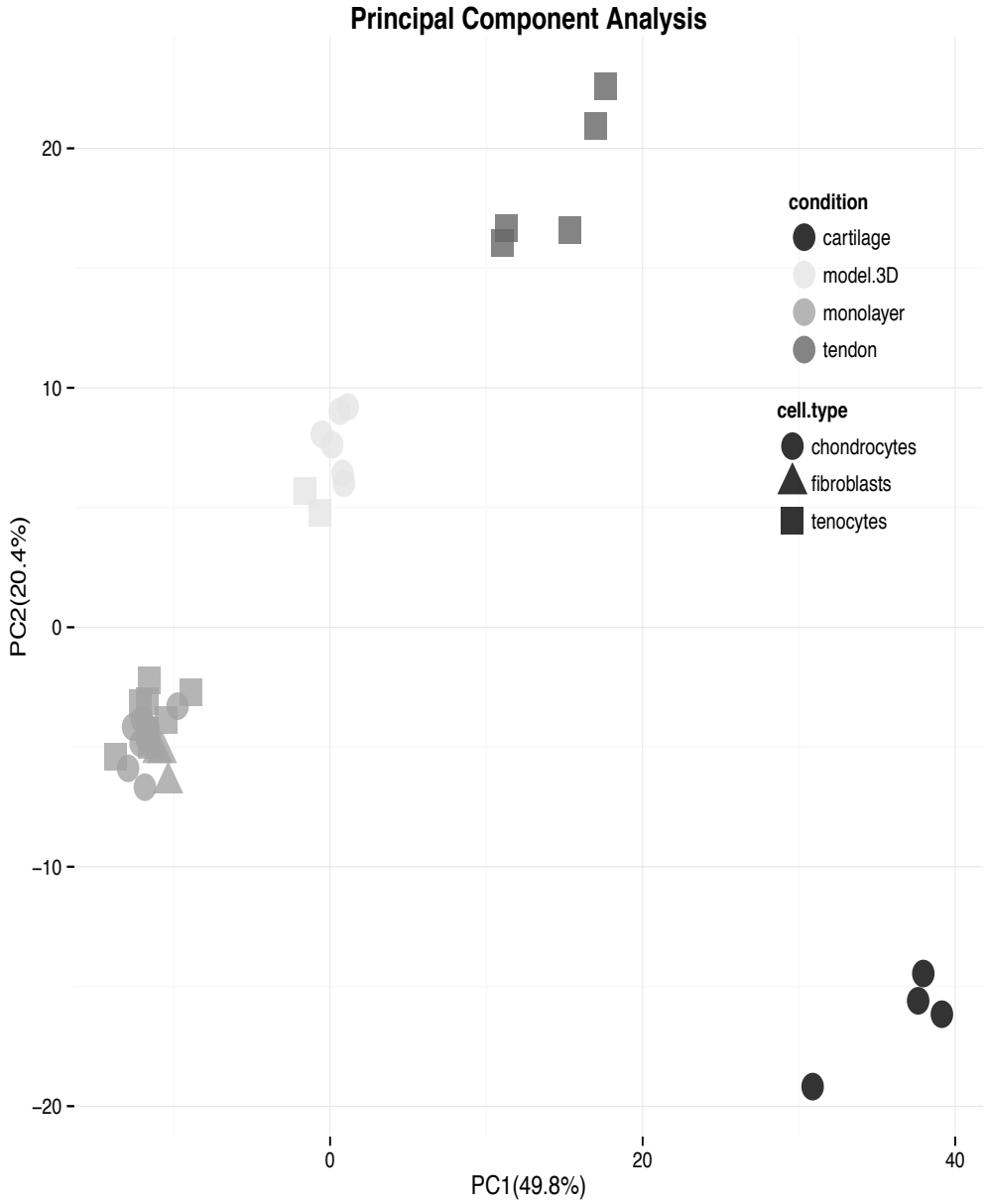
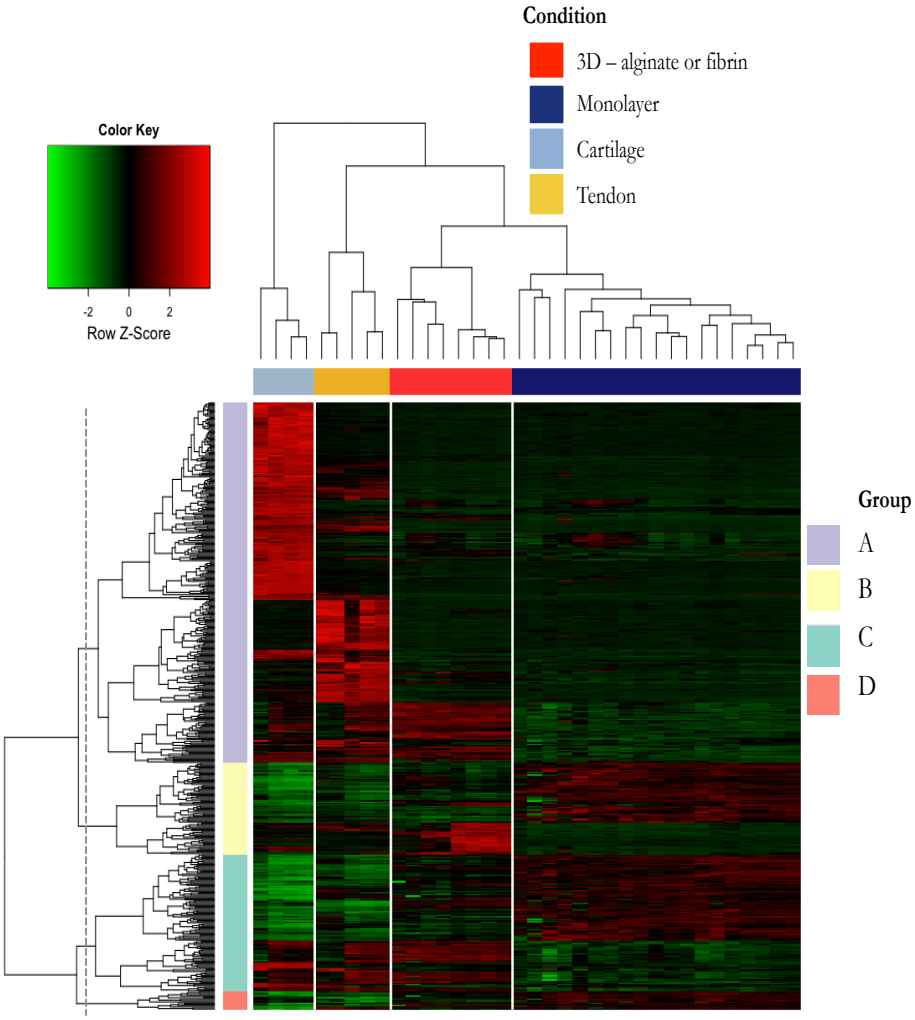


Figure 2.3: Principal component analysis of gene expression data from 36 Illumina arrays derived from three cell types (**Cell type**) and isolated from three environmental conditions (**Condition**) – native (cartilage or tendon matrix), monolayer (passage 5, dedifferentiated), or 3D (alginate or fibrin cultures). Plot presents the separation of samples based upon the first two principal components (PC1, PC2), which together explain >70% of the variation of the data in the top 500 most covariant genes. Although cartilage and tendon were highly divergent it was not possible to discriminate between cell types in monolayer and three-dimensional cultures, i.e. the plot reflects variation relating to the environmental conditions.



Group annotations based on defined clades

A: Contractile fibre | muscle system process | actin binding | cytoskeletal protein biding | actin filament-based process | immune system process | defense response | extra-cellular region | organ development | system development

B: Extracellular space | Response to external stimulus | Response to wounding | inflammatory response | developmental response | cell differentiation | cell migration

C: Translational elongation | ribonucleoprotein complex | structural constituent of ribosome | cytoplasm | extracellular matrix

D: Lysosome* | chemical homeostasis* | extra-cellular matrix*

Figure 2.4: Heatmap represents a matrix of scaled gene expression values (rows) for the 500 most covariant genes across 36 samples (columns), see figure legends. Dashed vertical line bisects clades into four groups (legend) for functional annotation – clades show general grouping with condition specific annotations in text and **SD2**. Gene ontology based upon significant terms (p<0.05) in each group except **D**, which is too small for significant annotation.

2.3.3: Differential gene expression

Overview

For each of the conditions considered in this study selected pairwise comparisons were undertaken to determine the number of differentially expressed genes in each, **Table 2.3**. The greatest number of differentially expressed genes were found between native tissues and their monolayer expanded equivalent, for example, a total of 2709 genes were up- and down-regulated between native cartilage and dedifferentiated chondrocytes. In comparison, 2352 genes were differentially expressed between native cartilage and chondrocytes in alginate, but only 289 between dedifferentiated tenocytes and chondrocytes after statistical threshold filtering. There were fewer differentially expressed genes between native tendon and tenocytes in monolayer or fibrin constructs than the equivalent comparisons for cartilage. The fewest statistically significant differentially expressed genes were found between monolayer-expanded tenocytes and fibroblasts; a total of 270 were either up- or down-regulated.

Comparison	n.C	n.T	d.C	d.T	d.F	Alginate	Fibrin
n.C		823	1503			1244	
n.T	1142			964			649
d.C	1206			154		522	
d.T		734	135		113		294
d.F				157			
Alginate	1108		900				158
Fibrin		580		627		265	

Higher expression

Lower expression

Table 2.3: Matrix of up- and down-regulated genes for selected pairwise comparisons involving different environmental conditions for chondrocytes and tenocytes. Values indicate the number of up- or down- regulated genes with a $\log_2$FC >+/- 0.5, FDR <0.01 and a log-odds ratio of expression >0. Duplicate Entrez gene identifiers are removed. Table Code: **n.C** – native cartilage; **n.T** – native tendon; **d.C** – dedifferentiated chondrocytes; **d.T** – dedifferentiated tenocytes; **d.F** – dedifferentiated fibroblasts; **Alginate** – chondrocytes in alginate beads; **Fibrin** – tenocytes in fibrin constructs. The fewest differentially expressed genes were found between cultured cells.

Analysis of differential gene expression by cell type

CHONDROCYTES

In dedifferentiated chondrocytes (monolayer) the most highly expressed genes were matrix metalloproteinase 3 (*Mmp3*), transforming growth-factor beta 2 (*Tgf-β2*) and thrombospondin 2 (*Thbs2*); compared to native cartilage the greatest reduction in expression is found in genes encoding heamaglobin alpha and beta chains, *Hbb* and *Hbb-b*, *Hba1,* defensin *Defa5* and the cathelicidin-related antimicrobial peptide *Cramp/Camp*. Monolayer chondrocytes also demonstrated higher expression of mesenchymal markers *Thy-1* and prion protein gene, *Prnp*, epithelial-mesenchymal transition regulator *Snai1* and bHLH transcription factor

Twist1. The TGF-β signalling inhibitor *Smad7*, was also higher in monolayer culture.

Native chondrocytes expressed higher levels of collagen type II (*Col2a1*), aggrecan (*Acan*), thrombospondin 4 (*Thbs4*), clusterin (*Clu*), dentin matrix acidic phosphoprotein (*Dmp1*), integrin-binding sialoprotein (*Ibsp*), and proteoglycan 2 and 3 (*Prg2/3*) than monolayer-expanded chondrocytes. Expression of CCN-family genes, connective tissue growth factor (*Ccn2/Ctgf*) and Wnt1-inducible signalling pathway proteins *Wisp1* and *Wisp2*, and Wnt-signalling gene frizzled family receptors *Fzd1*, *Fzd2* and *Fzd8* were all lower in native cartilage than monolayer.

There was notable differential expression of homeobox genes across native cartilage and monolayer. Monolayer was associated with higher expression of *Pitx1*, a hind-limb coding gene, and *Prrx2*, a differentiation-associated homeobox gene. In native cartilage HOP homeobox (*Hopx*) and SATB homeobox 2 (*Satb2*) were more highly expressed.

Relative to native cartilage, chondrocytes in alginate culture expressed the chemokine ligand 1 *Cxcl1* most highly, followed by prostaglandin D2 synthase (*Ptgds*) and *Mmp3*. The FBJ osteosarcoma oncogene *Fos, Ccn3* (formerly nephroblastoma-overexpressed, *Nov*) and chitinase-3 like-1 (*Chi3l1*) were also more highly expressed in alginate beads relative to native cartilage. In comparison to monolayer chondrocytes, alginate bead cultures expressed higher levels of interleukin 6 (*Il-6*), alarmin genes (*S100a4*, *S100b*) and prostaglandin-endoperoxide synthase 2 (*Ptgs2/Cox2*). A number of chondrogenesis-associated genes were also

more highly expressed, *Scrg1* (stimulator of chondrogenesis 1), inhibitor of DNA binding 1 (*Id1*), *Bhlhb2* (basic helix loop helix transcription factor *Dec1*).

TENOCYTES

The most highly expressed genes in dedifferentiated tenocytes included phospholipase A2 (*Pla2g7*), integrin-α11 (*Itga11*), heme oxygenase 1 (*Homx1*) and secreted phosphoprotein 1 (*Spp1*). The serine protease *Serpine1* and the transcription factor *Cebpb* were also more highly expressed in monolayer relative to native tendon. Monolayer tenocytes, in addition to expressing genes in common with chondrocytes (*Thy1*, *Prnp*, *Twist1*) also expressed elevated levels of biglycan (*Bgn*) and the homolog of *slit* (*Drosophila*), *Slit3*.

Those genes most highly expressed in native tendon included troponins and myosins. Specifically troponin I type 2 (*Tnni2*), actin α1 (*Acta1*) and creatine kinase (*Ckm*) had higher expression in native tendon relative to monolayer. The differentiation marker tenomodulin (*Tnmd*) and tendon-associated gene Mustang (*Mustn1*) had significantly lower expression in monolayer cultures. Elastin (*Eln*), keratocan (*Kera*), lubricin (*Prg4*), dermatopontin (*Dpt*), apolipoprotein (*ApoE*) and bone morphogenetic protein encoding genes *Bmp1* and *Bmp7* were more highly expressed in native tendon than in monolayer.

When compared to native tendon, tenocytes in fibrin cultures expressed higher levels of metallothionein 1a (*Mt1a*), the BMP-antagonist gremlin 1 (*Grem1*) and enolase 2 (*Eno2*), a neuron-associated enolase isoenzyme. The tenasin N/W isoform, *Tnn*, was highly expressed in fibrin cultures. In addition to the chemokines found in alginate beads (*Cxcl1*, *Ptgs2*), hypoxia-inducible factor 1 (*Hif1a*), RUNT transcription factor *Runx1*, and *Gpnmb*/osteoactivin were also

highly expressed. Relative to alginate beads, fibrin cultures exhibited higher expression of microfibrillar-associated protein 5 (*Mfap5/Magp2*), thrombospondin 4 (*Thbs4*), and the *Meox2* homeobox gene.

Consensus genes by condition

Principal component and differential expression analysis indicated that a number of genes were commonly expressed in native tissue, or monolayer and three-dimensional cultures derived from either cartilage or tendon. To describe common functions the consensus genes for different conditions were identified, **Figure 2.5**.

NATIVE TISSUE TO MONOLAYER CULTURES

There were 861 genes that were differentially expressed in both cartilage and tendon transitions to monolayer culture. The top differentially expressed genes unique to tenocytes or chondrocytes are presented in **Table 2.4a**. Of the consensus genes for cartilage and tendon only 34 of 861 differed in the direction of the fold change. Genes found to be differentially expressed in both tissues included lubricin (*Prg4*), the HOP homeobox gen (*Hopx*), osteocalcin (*Bglap*) and the troponin I, type 2 (*Tnni2*). The top five highest-ranked consensus genes for native cartilage and tendon were: *Hba1, Hbb,* transferrin (*Tf*)*, Hbb-b1,* complement factor D (*Cfd*) and major histocompatibility complex gene *RT1-Da.* Relative to monolayer culture 457 genes with lower expression in both cartilage and tendon with the highest ranking genes including: transgelin (*Tagln),* *Had11b1* (hydroxysteroid 11-beta dehydrogenase 1)*,* collagen type VIII, alpha 1 subunit (*Col8a1*)*, C1qtnf5* (C1q and tumor necrosis factor related protein 5)*, Itga11* and *Serpine1*.

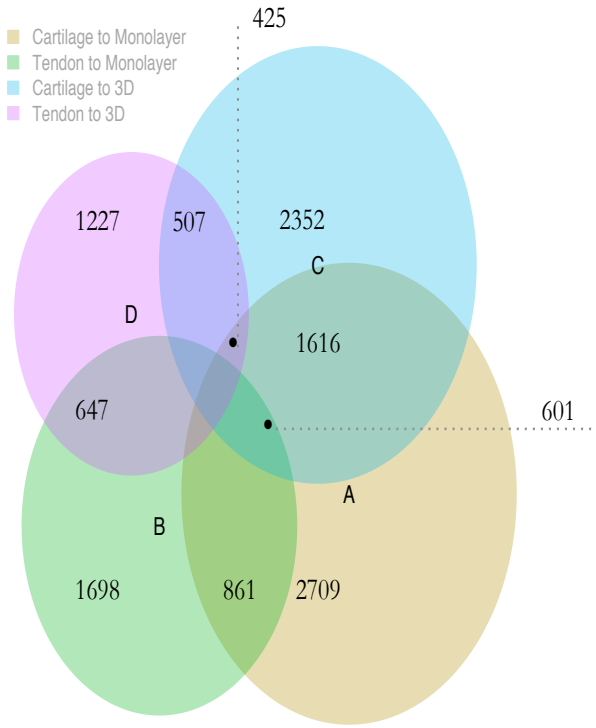


Figure 2.5: Euler diagram indicating the proportion of overlap between different pair-wise comparisons. All comparisons are made with respect to the native tissue and duplicate Entrez gene entries are removed. For example, in the union between genes differentially expressed in cartilage vs. monolayer chondrocytes and tendon vs. monolayer tenocytes, 861 genes are differentially expressed in both comparisons. **A**: native cartilage vs. chondrocytes; **B**: native tendon vs. tenocytes; **C**, native cartilage vs. alginate beads; **D**, native tendon vs. fibrin constructs.

NATIVE TISSUE TO THREE-DIMENSIONAL CULTURES

When differentially expressed genes in three-dimensional cultures are considered, **Table 2.4b**, 283 genes were found to be down-regulated in native tissue relative to both alginate and fibrin cultures. In three-dimensional cultures AP-1 components *Fos* and *Jun*, the transmembrane glycoprotein osteoactivin gene *Gpnmb*, clusterin (*Clu*) and the bone morphogenetic protein receptor, type 1a (*Bmpr1a*) were all more highly expressed relative to native tissue. Top five ranked consensus genes in three-dimensional cultures were *Errfi1* (ERBB receptor feedback inhibitor 1), *Rasd1* (dexamethasone-induced Ras-related protein 1) and the serine peptidase inhibitor *Serpina3n*. Consensus gene lists are provided in **SD2.13-2.14**. Genes found to be uniquely expressed in each condition are provided in **SD2.15-SD2.18**.

Table 2.4a: Differentially expressed genes and consensus comparisons described in **Figure 2.5** considered in more detail for the transition from the native to monolayer condition. Top genes unique to each pairwise comparison are presented with genes showing higher expression presented in the top frames and those showing lower expression in the bottom frames. Consensus genes are presented as a prioritised gene list based upon a rank aggregation method using the adjusted p-value. Standard gene symbols are presented, but not italicised.

Gene Symbol	log_2FC	FDR
Myl1	7.3	9.8e-19
Tnnt3	6.8	3.8e-18
Mb	5.6	6.5e-21
Ttn	5.4	3.0e-12
Hfe2	5.3	1.8e-15
Pygm	5.2	3.5e-15
Des	5.2	4.1e-14
Cox8b	5.2	3.2e-20
Myot	4.9	2.7e-10
Pvalb	4.8	6.1e-12
Myh7	4.8	1.1e-10
Tnmd	4.7	1.1e-08

Higher expression

Gene Symbol	Score
Hba1	6.7e-05
Hbb	1.3e-04
Tf	4.6e-04
Hbb-b1	8.7e-03
Cfd	1.1e-03
RT1-DA	1.3e-03
Prg4	1.3e-03
Aif1	2.1e-03
Pla2g2a	2.1e-03
RGD1562323	2.3e-03

Ranked consensus native tissue, up-regulated genes

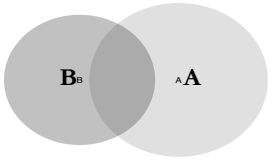
Gene Symbol	log_2FC	FDR
Np4	7.3	9.8e-19
Car2	6.8	3.8e-18
Prg2	5.6	6.5e-21
Ifitm6	5.4	3.0e-12
RGD1565970	5.3	1.8e-15
Retnlg	5.2	3.5e-15
Dmp1	5.2	4.1e-14
RGD1561694	5.2	3.2e-20
Ca1	4.9	2.7e-10
Ctsg	4.8	6.1e-12
Serpinb1a	4.8	1.1e-10
Mpo	4.7	1.1e-08

Higher expression

Unique native tendon vs. dedifferentiated **tenocytes**

Gene Symbol	log_2FC	FDR
Pla2g7	-3.6	1.3e-09
Spp1	-3.4	5.5e-06
Prss35	-3.2	2.2e-08
Gadd45a	-2.9	8.4e-08
Fam198b	-2.8	1.1e-05
Mgst2	-2.7	3.3e-08
Mta1a	-2.6	1.3e-05
Clec2d	-2.5	7.4e-10
Col5a2	-2.3	1.0e-04
Ddit3	-2.2	2.4e-05
Smoc2	-2.2	3.9e-05
Plod2	-2.03	8.8e-05

Lower expression



Ranked consensus monolayer, down-regulated genes

Gene Symbol	Score
Tagln	3.9e-04
Hsd11b1	4.3e-04
Col8a1	4.6e-04
Itga11	7.0e-04
C1qtnf5	8.7e-04
Serpine1	1.1e-03
Gpr176	1.6e-03
Nqo1	1.7e-03
Insig1	1.9e-03
Plod1	2.3e-03

Lower expression

Unique native cartilage vs. dedifferentiated **chondrocytes**

Gene Symbol	log_2FC	FDR
Mmp3	-5.4	4.5e-07
Abi3bp	-4.4	1.5e-10
Prss23	-4.1	9.9e-08
Nov	-3.8	1.0e-06
Loxl2	-3.7	3.1e-11
Bag3	-3.6	1.2e-10
Pmp22	-3.5	9.6e-16
Timp2	-3.5	6.3e-05
Smoc1	-3.4	1.1e-13
Aox1	-3.3	3.6e-08
Nbl1	-3.3	1.8e-11
Lgals1	-3.2	4.6e-06

Table 2.4b: Differentially expressed genes and consensus comparisons described in **Figure 2.5** considered in more detail for the transition from the native to three-dimensional culture condition. Top genes unique to each pairwise comparison are presented with genes showing higher expression presented in the top frames and those showing lower expression in the bottom frames. Consensus genes are presented as a prioritised gene list based upon a rank aggregation method using the adjusted p-value. Standard gene symbols are presented, but not italicised.

Unique native tendon vs. fibrin constructs

Gene Symbol	log_2FC	FDR
Acta1	8.6	1.3e-10
Tnnc2	8.2	8.0e-15
Mylpf	7.3	1.6e-15
Myl1	7.3	4.2e-14
Tpm1	6.9	1.4e-11
Tnnt3	6.4	5.3e-13
Myoc	6.3	3.3e-19
Mb	5.7	1.4e-16
Tnmd	5.6	1.1e-06
Hfe2	5.3	3.1e-11
Pygm	5.2	6.9e-11
Ttn	5.2	4.0e-07

Higher expression

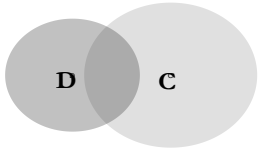
Ranked consensus native tissue, up-regulated genes

Gene Symbol	Score
Tf	1.6e-3
Fcer1g	2.2e-3
Hba1	2.2e-3
Hbb	25.5e-3
Pf4	8.7e-3
RT1-Da	0.015
Hbb-b1	0.018
Coro1a	0.019
Lyl1	0.026
C1qc	0.031

Higher expression

Unique native cartilage vs. alginate cultures

Gene Symbol	log_2FC	FDR
Defa5	8.1	1.5e-22
RatNP-3b	7.3	1.1e-23
Bglap	7.2	1.0e-11
Np4	6.7	1.6e-25
Car2	6.4	7.1e-28
Prg2	6.1	1.6e-26
Ifitm6	5.9	2.4e-30
Plac8	5.8	4.8e-25
Col2a1	5.7	7.5e-08
Retnlg	5.3	5.8e-23
Dmp1	5.3	2.3e-23
Arhgdib	5.2	8.6e-26



Gene Symbol	log_2FC	FDR
Mt1a	-4.4	4.2e-07
Grem1	-4.3	2.7e-09
Tnn	-3.5	4.3e-06
Npy	-3.3	2.2e-03
Prdx5	-3.2	5.4e-06
Car9	-3.2	4.9e-07
Mgst1	-3.2	1.2e-4
Pla2g7	-2.8	2.6e-06
Cthrc1	-2.8	8.9e-05
LOC687649	-2.7	2.5e-11
Enpp3	-2.6	7.7e-05
Mmp11	-2.6	3.9e-05

Lower expression

Ranked consensus 3D cultures, down-regulated genes

Gene Symbol	Score
Errfi1	5.5e-4
Rasd1	1.2e-3
Cp	1.7e-3
Serpin3n	2.8e-3
Usmg5	0.0107
Clu	0.011
Cfb	0.012
LOC100364062	0.016
Mtx1	0.02
Rbbp7	0.03

Lower expression

Gene Symbol	log_2FC	FDR
Ptgds	-5.3	4.7e-22
MMP3	-5.2	9.1e-07
Hsd11b1	-5.2	4.6e-12
Aox1	-4.7	2.7e-11
Nov	-4.4	1.1e-07
Mlana	-4.2	1.2e-10
Nbl1	-4.2	4.1e-14
Fbln5	-3.8	1.0e-10
Ier3	-3.7	7.2e-17
Insig1	-3.7	1.7e-09
Atf3	-3.6	1.5e-24
Gdf15	-3.5	1.3e-16

2.3.4: Gene Ontology Functional Annotation

NATIVE TISSUE

All terms reported are significantly enriched ($p<0.001$) after adjustment for multiple testing. The genes most highly expressed in native cartilage were significantly associated with the gene ontology biological process terms 'immune system process', 'immune response', 'cell cycle' and 'defense response' using a hypergeometric analysis.

For native tendon the most highly enriched terms were those relating to 'muscle system process', 'muscle contraction', and 'immune system process'. The gene expression profile was also associated with 'skeletal muscle tissue development' and 'actin filament-based process'. A full list is available in **SD2.19-SD2.24**.

MONOLAYER

Dedifferentiated (monolayer) chondrocytes and tenocytes demonstrated some overlap in the biological process gene ontology terms that were significantly enriched. The terms 'single-organism metabolic process', 'sterol biosynthetic process' and 'small molecule metabolic process' were associated with genes more highly expressed in dedifferentiated cells. The gene expression profile in monolayer cells was described by development associated terms 'anatomical structure development' and 'developmental process'; terms associated with 'extracellular matrix organisation' and 'cell substrate adhesion' were also enriched for both cell types. In tenocytes 'cell redox homeostasis' was the most highly enriched term; in chondrocytes 'oxidation-reduction process' was enriched.

THREE-DIMENSIONAL CULTURE

In the three-dimensional culture context genes more highly expressed in alginate beads containing chondrocytes were associated with biological processes terms related to: 'single organism metabolic process', 'oxidation-reduction process', 'oxoacid metabolic process', 'response to oxidative stress' and 'lipid metabolic process'. The gene expression profile was also described by developmental terms relating to 'vasculature development', 'cardiovascular development', 'post-embryonic development' and 'regulation of anatomical structure morphogenesis'. Fibrin constructs containing tenocytes were enriched for biological process terms associated with 'translation', 'metabolic process', 'collagen catabolic process' and 'single-organism metabolic process'. The terms 'response to oxidative stress' and 'apoptotic signaling pathway' are common to both fibrin and alginate cultures.

METABOLIC FUNCTIONS AND CELLULAR COMPARTMENTS

To define the common metabolic functions and cellular locations of differentially expressed gene in different conditions functional annotations for 'metabolic function' and 'cellular compartments' for consensus genes were analysed.

For consensus genes derived from the native tissue to monolayer comparison 'actin filament binding', 'integrin binding' and 'cytoskeletal protein binding' were enriched. For native to three-dimensional culture comparison 'lipid particle binding', 'glycosaminoglycan binding' and 'carbohydrate binding' were enriched metabolic functions. The term 'oxidoreductase activity' was common to both analyses. The cellular compartment annotations was described by the terms: 'cytoplasmic part', 'membrane-bound organelle' and 'cytoplasm' in both analyses, **SD2.23-SD2.24**.

Differentially expressed genes were also functionally annotated using the Ingenuity® Pathway Analysis knowledge base to include further disease and physiology annotations. The summary annotations for the three comparisons, native – monolayer – 3D, for cartilage (**Table 2.5a**) and tendon (**Table 2.5b**) are presented.

Comparison	IPA descriptors
Native cartilage to monolayer 	**DD**: Inflammatory response | Connective Tissue Disorders | Skeletal muscle disorders. **MCF:** Cell growth and proliferation | Cell movement | Cell death and survival **PS**: Immune Cell Trafficking | Hematological system development | Tissue morphology | Tissue development
Monolayer to alginate 	**DD**: Dermatological disease and conditions | Cancer | Cardiovascular disease | Neurological disease | Skeletal and muscular disorders **MCF**: Cell death and survival | Cell growth and proliferation | Cellular movement **PS**: Immune cell trafficking | Hematological system development and function | Cardiovascular system development and function | Organismal development
Alginate to Native cartilage 	**DD**: Inflammatory response | Connective tissue disorders | Inflammatory disease | Skeletal muscle disorders | Immunological disease; **MCF**: Cellular movement | Cellular growth and proliferation | Cell death and survival | Cell-to-cell signalling and interaction **PS**: Immune cell trafficking | Hematological system development and function | Tissue morphology | Tissue development | Cardiovascular system development and function

Table 2.5a: Functional descriptors of differentially expressed gene lists derived using Ingenuity Pathway Analysis for chondrocyte comparisons. Legend: **DD** – Diseases and Disorders; **MCF**: Molecular and Cellular Function; **PS**: Physiological System Development and Function. These give another level of analysis of the data; within each parent term lie numerous 'child' terms carrying their own *p*-values and activation *z*-scores. The comparison of alginate to native cartilage uses three-dimensional cultures as the baseline condition consistent with pathway topology analysis figures.

Comparison	IPA descriptors
Native tendon to monolayer 	**DD**: Neurological disease | Psychological disorders | Skeletal and muscular disorders | Hereditary disorder | Organismal injury and abnormalities **MCF:** Cell growth and proliferation | Cell death and survival | Cell morphology | Cellular movement **PS**: Organ morphology | Skeletal and muscular system development and function | Cardiovascular system development and function | Tissue morphology | Embryonic development
Monolayer to fibrin 	**DD**: Dermatological disease and conditions | Neurological disease| Connective tissue disorders | Hereditary disorder | Opthalmic disease **MCF**: Protein synthesis | Cell death and survival | Cell growth and proliferation | Cellular movement | Cellular development **PS**: Skeletal and muscular system development | Cardiovascular system development and function | Organismal development | Tissue development | Connective tissue development and function
Fibrin to native tendon 	**DD**: Neurological disease | Cardiovascular disease | Skeletal and muscular disorders | Psychological disorders | Hereditary disorder **MCF**: Cellular movement | Cellular growth and proliferation | Cell death and survival | Cell-to-cell signalling and interaction **PS**: Organ morphology | Skeletal and muscular system development and function | Cardiovascular system development and function | Embryonic development | Organ development

Table 2.5b: Functional descriptors of differentially expressed gene lists derived using Ingenuity Pathway Analysis for tenocyte comparisons. Legend: **DD** – Diseases and Disorders; **MCF**: Molecular and Cellular Function; **PS**: Physiological System Development and Function. These give another level of analysis of the data; within each parent term lie numerous 'child' terms carrying their own *p*-values and activation z-scores. The comparison of fibrin cultures to native tendon uses three-dimensional cultures as the baseline condition consistent with pathway topology analysis figures.

2.3.5: Pathway topology - prediction of perturbed pathways

A pathway topology technique was employed (SPIA), which makes use of both differentially expressed gene effect size and pathway topology to predict perturbations of KEGG canonical signalling pathways and predicts an activation state. Both positive and negative $\log_2$-fold-changes from comparisons were used as inputs to SPIA to provide an adequate description of each transition. Ingenuity® Pathway Analysis canonical pathways, derived from over-representation analysis methods, were also assessed. The most perturbed pathways are provided in **SD2.25-2.26** with annotated HTML links to KEGG pathways.

Native to monolayer transition

Consistently the most significantly perturbed pathways for the native to monolayer transition for cartilage were 'cell cycle', 'systemic lupus erythematosus' and 'chemokine signalling pathway', **Figure 2.6**. For both cartilage and tendon native to monolayer comparisons the canonical pathways 'rheumatoid arthritis' and 'systemic lupus erythematosus' were predicted to be activated. Specifically in tendon to monolayer 'focal adhesion' and 'complement and coagulation cascades' were predicted to be activated, **Figure 2.7**, whilst the canonical pathway 'Parkinson's disease' was predicted to be inhibited.

Monolayer to three-dimensional culture transition

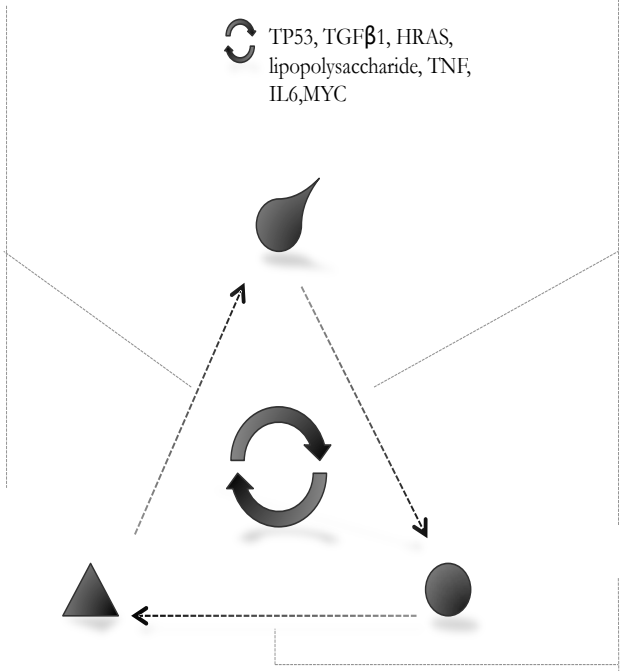
In the transition from monolayer to alginate cultures chondrocytes were predicted to have activation of the pathways 'focal adhesion', 'PI-3K/Akt signalling pathway' and 'ECM-receptor interaction; 'chemokine signalling pathway' was the most significant pathway and this was predicted to be inhibited; this was also true

for the monolayer tenocyte to fibrin culture transition. In the latter comparison the 'HIF1-signalling pathway' was predicted to be activated.

| IPA
- Acute phase response signalling;
- LXR/RXR activation;
- Complement system
- IL-17a signalling in fibroblasts

| KEGG
+ Focal Adhesion
+ PI3K-Akt signalling pathway
+ ECM-receptor interaction

- Chemokine signalling pathway
- Huntington's disease
- Alzheimer's disease



| IPA
- Mismatch repair in eukaryotes
- Heme biosynthesis II
- NRF2-mediated oxidative stress response
- Granulocyte adhesion and diapedesis

| KEGG
+ Cell cycle;
+ Osteoclast differentiation;
+ TNF signalling pathway.
- Systemic Lupus erythematosus ;
- Rheumatoid arthritis;
- PPAR signalling pathway.

| IPA
- Integrin signalling
- Atherosclerosis signalling
- Leukocyte extravasation signalling

| KEGG
+ Systemic Lupus erythematosus
+ Chemokine signalling pathway
+ Rheumatoid arthritis
- Cell cycle
- Transcriptional misregulation in cancer
- HIF-1 signalling pathway

Legend

Native chondrocytes

Monolayer chondrocytes

3D culture chondrocytes

Pairwise comparison

Common upstream regulators

Figure 2.6: Schematic Diagram: Using genes found to be differentially expressed in three pairwise comparisons, and their associated effect size (log_2 fold-change), canonical KEGG pathway activation was inferred using a pathway topology approach. Pathways are predicted to be activated (+) or inhibited (-) based upon global perturbation score and FDR-adjusted *p*-value. Arrows indicate the direction of the pairwise comparison. Predicted pathways from Ingenuity Pathway Analysis (IPA) are also shown, but do not use the pathway topology approach.

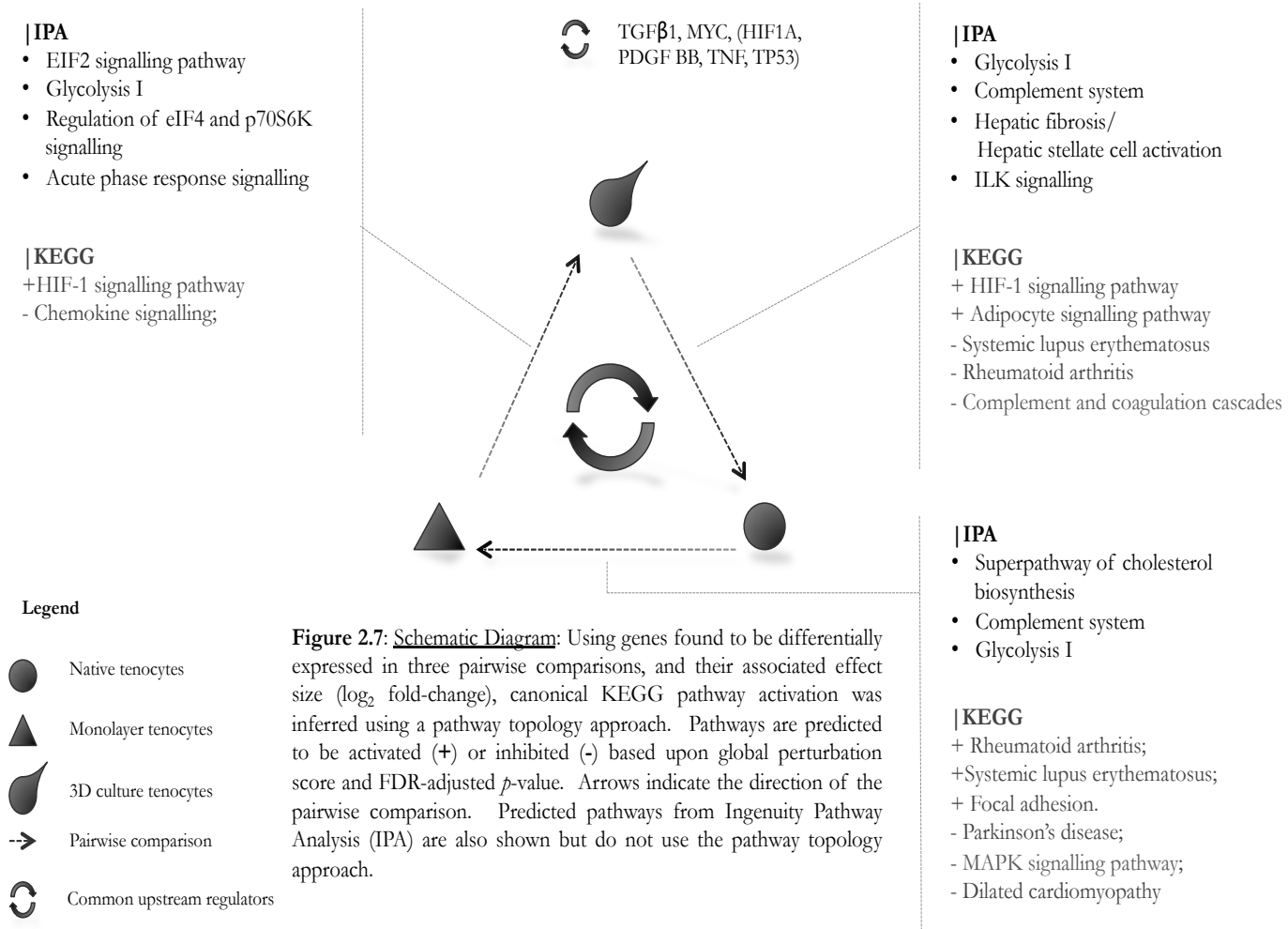


Figure 2.7: Schematic Diagram: Using genes found to be differentially expressed in three pairwise comparisons, and their associated effect size (log_2 fold-change), canonical KEGG pathway activation was inferred using a pathway topology approach. Pathways are predicted to be activated (+) or inhibited (-) based upon global perturbation score and FDR-adjusted *p*-value. Arrows indicate the direction of the pairwise comparison. Predicted pathways from Ingenuity Pathway Analysis (IPA) are also shown but do not use the pathway topology approach.

2.3.6: Inference of Upstream Regulators from Gene Expression Data

Upstream regulators of genes defined as differentially expressed in comparisons were inferred using Ingenuity® Pathway Analysis (IPA).

Top Scoring Upstream Regulators

The top scoring upstream regulators were ordered by overlap *p*-value. In the cartilage to monolayer comparison the top upstream regulators were inferred to be *Tp53*, *Tgf-β1*, lipopolysaccharide (LPS), *Hras*, and *Myc* (p < 1.86e-18). Of the two mechanistic networks generated from the top scoring regulators (z-scores) *Tp53* (-3.68, p = 9.01e-32, inhibited) contained 14 regulators with down-stream effects on 435 dataset genes; *Tgf-β1* (-2.6, p = 2.04e-26 inhibited) targeted 23 regulators with effects on 679 downstream genes represented in the differential gene expression analysis, **Figure 2.8**. For the *Tgf-β1* mechanistic network downstream targets of the root node included: *Smad7*, *Smad3*, *Il-6*, *Jun*, *Fos*, *Cebpb*, and *Nf-κbia*. Of these *Smad7* expression was down-regulated in dedifferentiated chondrocytes in culture.

For native tendon compared to monolayer-expanded tenocytes the top scoring upstream regulators by overlap p-value were *Tgf-β1*, *Hras*, dexamethasone, *Myc* and *Kras*. Only *Kras* (-2.3, p=1.2e-16, inhibited) had a predicted activation score above the threshold set by IPA and was predicted to be inhibited, however a *Tgf-β1* network was also predicted to be inhibited (-1.04, p=4e-23) as for chondrocytes. Only *Myod1* was within the differential expression lists ($\log_2$FC=1.67) and predicted to be activated (2.45, p=1.04e-12). This regulator was upstream of 14 core regulators and effected 237 genes within the expression dataset, **Figure 2.9**.

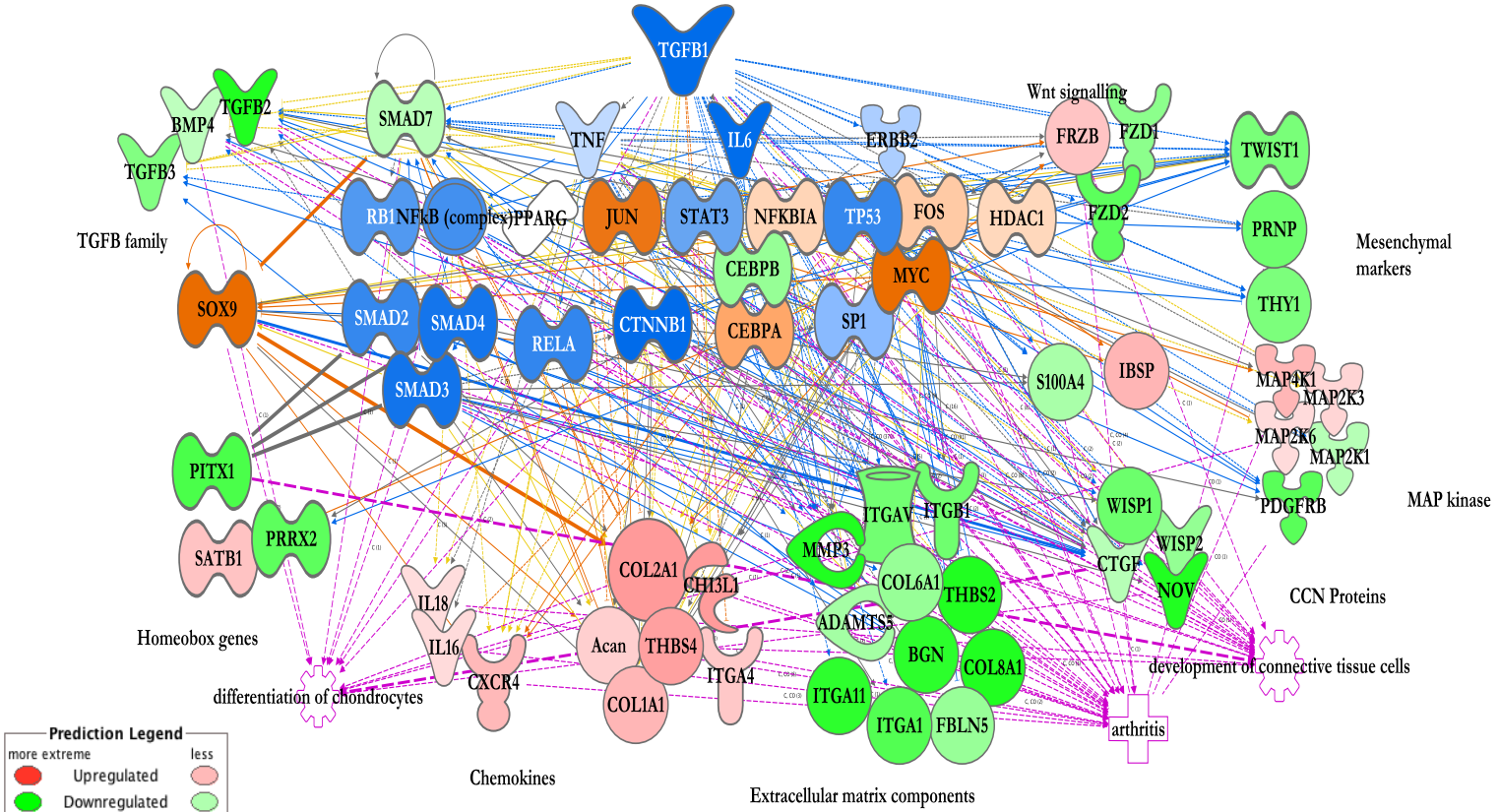


Figure 2.8: Ingenuity mechanistic network formed from Illumina data for **native cartilage to monolayer** transition. Up-stream master regulator predicted as *Tgf-β1* with downstream targets predicted from Ingenuity knowledge base and differentially expressed genes (**figure legend**) – genes more highly expressed in native chondrocytes are shown in graduate red, those down-regulated in monolayer culture are shown in graduated green. Genes are grouped into general functional groups including extra-cellular matrix. Genes in the mechanistic network are annotated with function and disease terms: 'differentiation of chondrocytes' (13 genes, p=1.5e-17); 'development of connective tissue cells' (19 genes, p=2.13e-21); 'arthritis' (32 genes, p=5.7e-22). *Smad7*, *Tgf-β1* -inhibitor, is highlighted to demonstrate a predicted mechanism for effects on *Col2a1* expression. Chondrogenesis regulator *Sox9* is predicted as activated in the network. Protein nomenclature used consistent with IPA output.

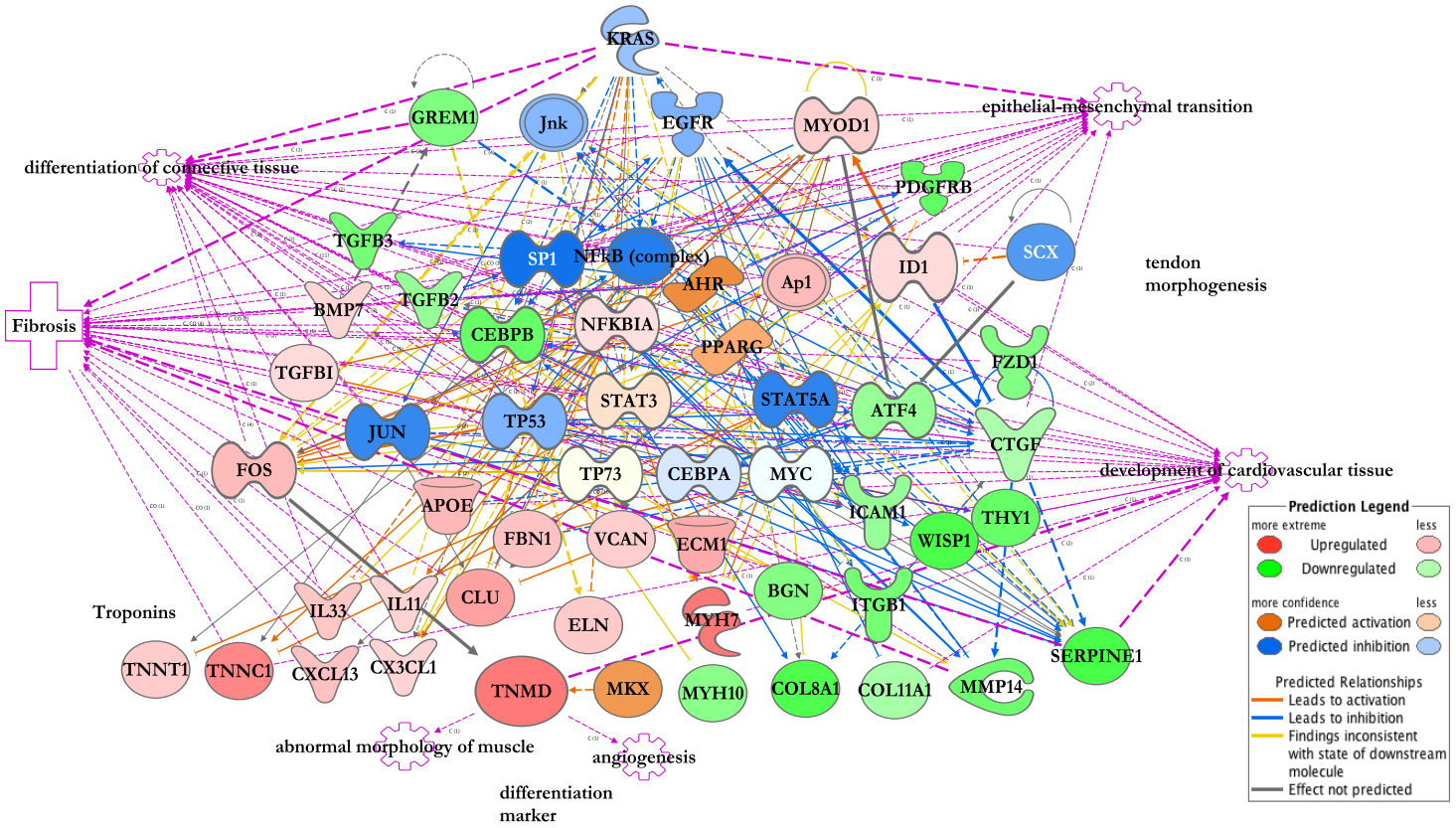


Figure 2.9: Ingenuity mechanistic network formed from Illumina data for **native tendon to monolayer** transition. Highest scoring up-stream master regulator predicted was *Kras* with downstream targets predicted from Ingenuity knowledge base and differentially expressed genes (**figure legend**). Genes in the mechanistic network are annotated with function and disease terms: 'differentiation of connective tissue' (29 genes, p=6.4e-25); 'epithelial-mesenchyme transition' (14 genes, p=2.03e-16); 'fibrosis' (27 genes, p=6.33e-26); 'development of cardiovascular tissue' (20 genes, 6.6e-17). Additional annotations given for tendon differentiation marker, tenomodulin (*Tnmd*) to highlight the lack of tendon-specific annotations. Key elements in the network include *Ccn2*/ connective tissue growth factor, gremlin 1 (*Grem1*), and inhibitor of DNA binding 1 (*Id1*, a bHLH transcription factor). Tendon morphogenesis factors scleraxis (*Scx*) and mohawk (*Mkx*) were added to the network provide tendon context. *Scx* is predicted to be inhibited in the native tendon context (the converse predicted for *Mkx*). Protein nomenclature used consistent with IPA output.

2.3.7: Differential expression validation: qPCR

Genes found to be differentially expressed in microarray analysis, or genes with known roles in cartilage and tendon morphogenesis were selected for qPCR validation.

Seven genes were selected for cartilage and chondrocyte gene expression validation by qPCR. These included: cartilage differentiation marker *Col2a1,* higher expression in native cartilage, hindlimb-patterning homeobox *Pitx1*, found to be more highly expressed in monolayer chondrocytes, and two homeobox genes found to be more highly expressed in native cartilage and tendon than monolayer, *Satb2* and *Hopx*. The chondrocyte development associated gene *Sox9*, predicted to be activated in cartilage by IPA, was selected, **Figure 2.8**. Two genes with mesenchyme-associated expression, found to be more highly expressed in monolayer, *Prnp* and *Thy1*, were also assessed.

For tenocytes and tendon the differentiation marker *Tnmd* was chosen for validation (more highly expressed in tendon than monolayer). Tendon-associated genes *Mkx* (differentiation) and *Scx* and *Mustn1* (development) were not differentially expressed, but were predicted by IPA to be activate and inhibited, respectively, in native tissue, **Figure 2.9.** The *Drosophila* tendon development-associated gene homolog *Slit3*, shown to be higher in monolayer tenocytes, was also selected.

There was general concordance in the direction and magnitude of expression changes between the three conditions for cartilage and tendon, although not all comparisons were found to be statistically significant. In summary, the hindlimb development homeobox gene *Pitx1* was significantly higher ($p<0.01$) in alginate

beads than either native or monolayer chondrocytes, **Figure 2.10A**. The cartilage development-associated homeobox gene *Satb2* was significantly more highly expressed in native cartilage than in either of the culture conditions ($p<0.05$). The corollary was true for *Prnp*, which was more highly expressed in culture conditions than in native cartilage ($p<0.05$), **Figure 2.10B**. Tenogenesis-associated *Mustn1* trended toward higher expression in native tendon than in monolayer cells ($p=0.057$), but the only significantly different was for tenomodulin (*Tnmd*) for native tendon compared to fibrin constructs ($p<0.05$), **Figure 2.11.** Significant differences in expression were not shown for *Slit3*, *Sox9* or *Thy-1* in tendon, monolayer or fibrin culture samples (not shown).

In summary, these findings suggest that two homeobox genes, *Pitx1* and *Satb2*, may represent useful differentiation markers in chondrocytes. Dedifferentiation was confirmed by the reduction in *Col2a1* and *Tnmd* expression in monolayer chondrocytes and tenocytes respectively, although this was only significant for the latter. Expression of developmental markers (*Scx*, *Sox9*) was equivocal in culture systems using qPCR.

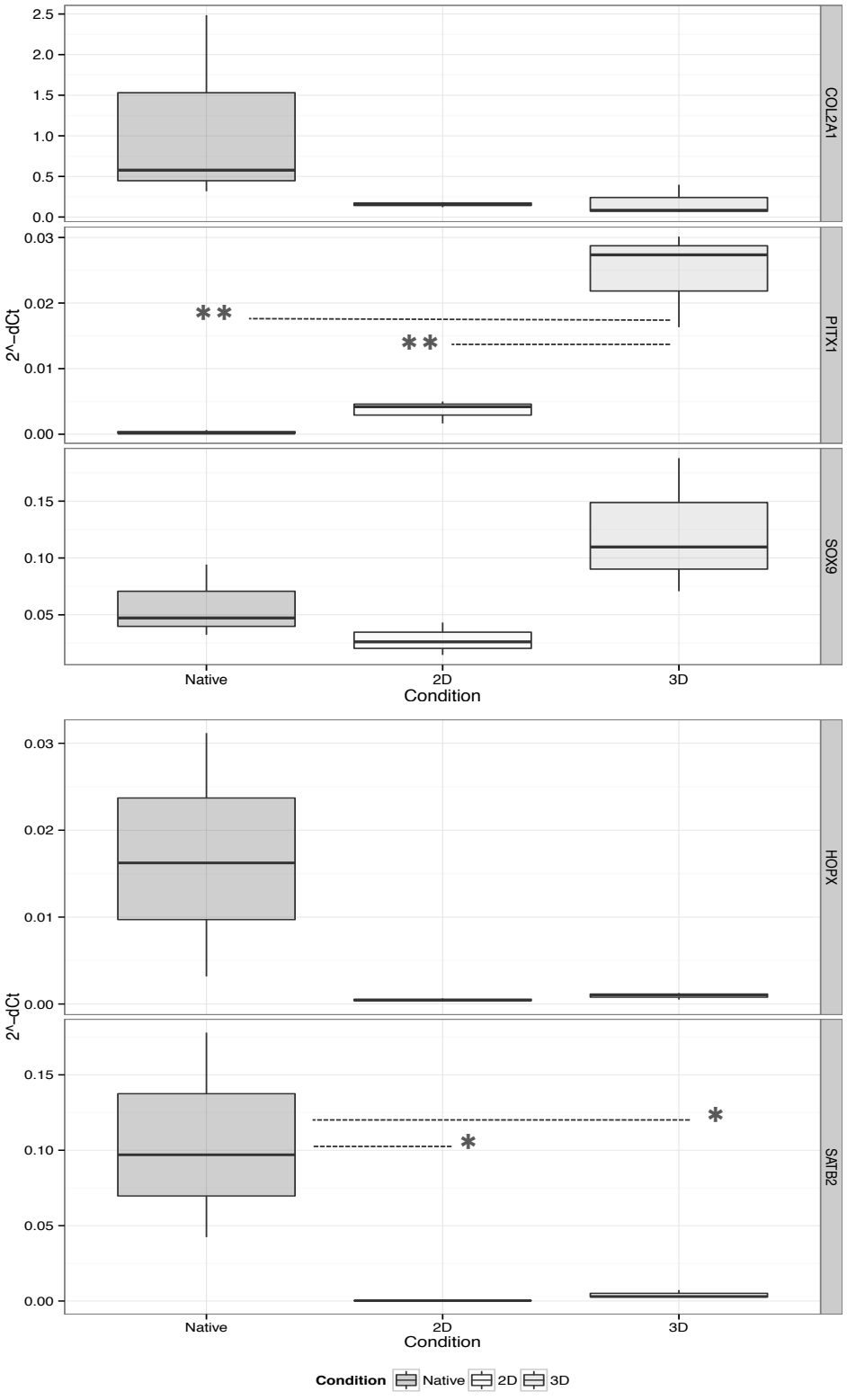
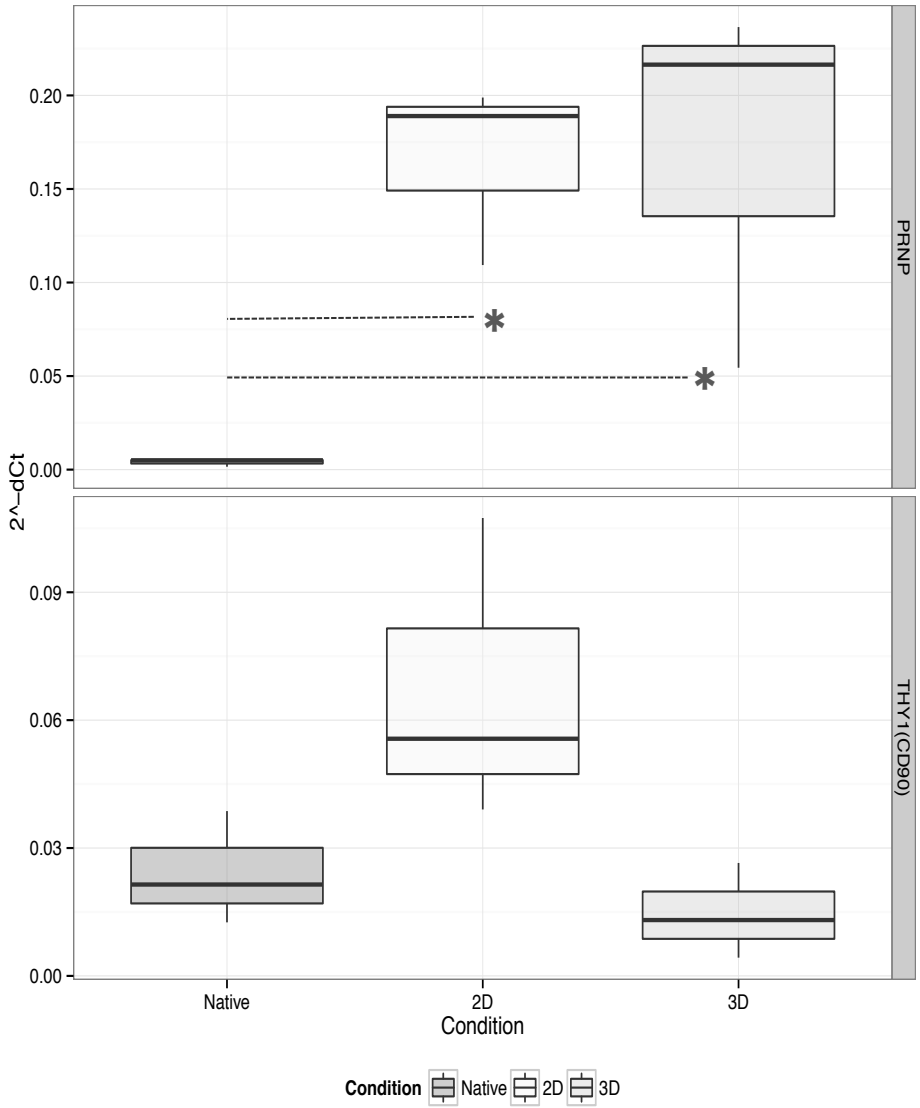


Figure 2.10A: Validation of expression changes in chondrocytes in three conditions (x-axis): **2D** – monolayer culture at passage 5; **3D** – alginate beads; **Native** – whole cartilage tissue. Boxplots present the distribution of the linear transformed C_t data (2^-dCt, y-axis). *Sox9*, a key regulator of chondrogenesis, was considered. *Sox9* expression is reduced in monolayer culture, but expression is higher in alginate cultures than in monolayer, but not significantly so. Significance code: * - $p<0.05$; ** - $p<0.01$; *** - $p<0.001$



2.10B*:* Validation of expression changes in chondrocytes in three conditions (x-axis): **2D** – monolayer culture at passage 5; **3D** – alginate beads; **Native** – whole cartilage tissue. Boxplots present the distribution of the linear transformed C_t data (2^-dCt, y-axis). Significance code: * - $p<0.05$; ** - $p<0.01$; *** - $p<0.001$. Mesenchymal stem cell markers *Prnp*, prion, and *Thy*-1 show higher expression in monolayer cultures; this is significant for *Prnp*. This analysis is performed using cartilage, monolayer chondrocytes and alginate beads.

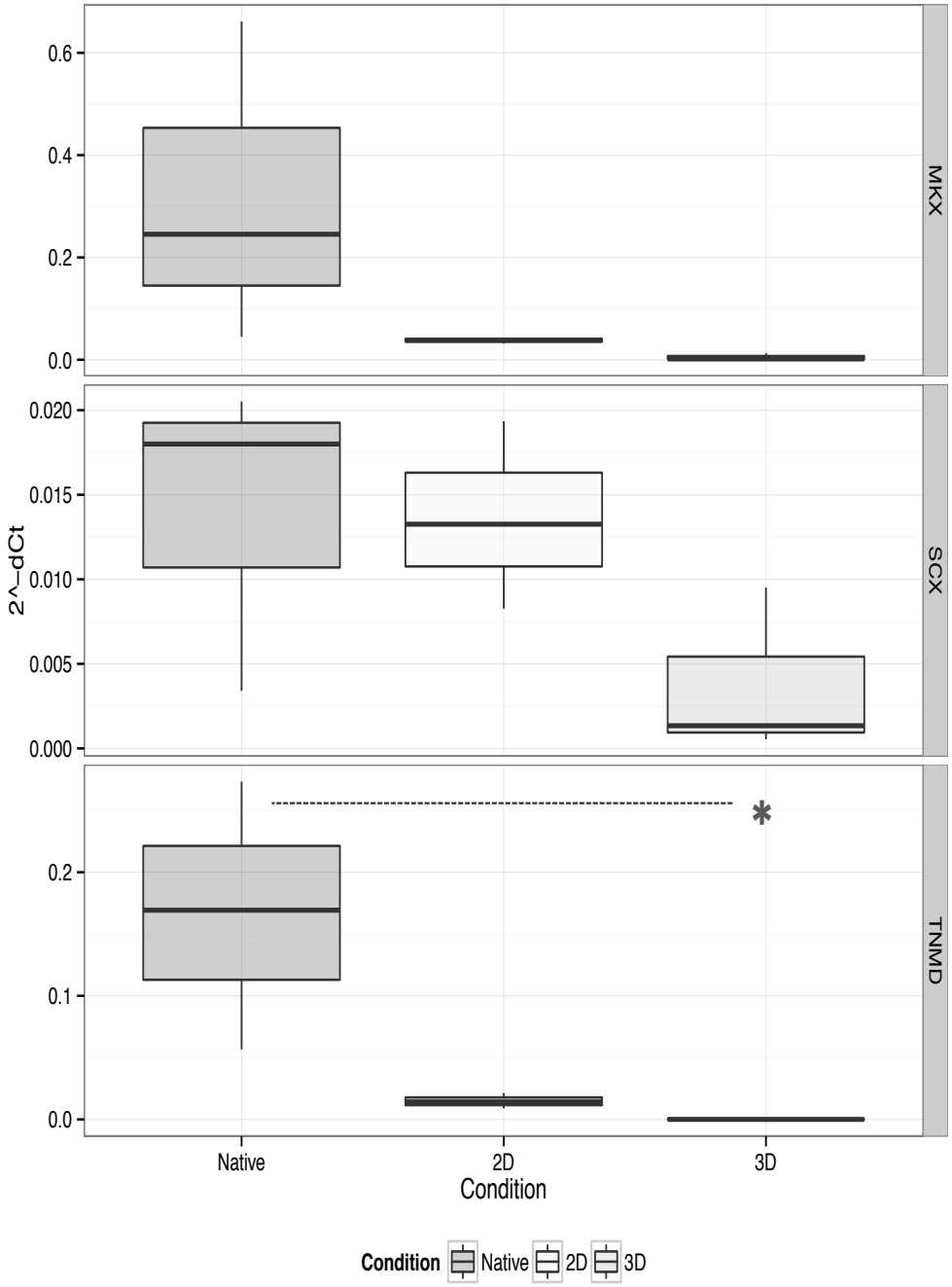


Figure 2.11: Validation of expression changes in tenoocytes in three conditions (x-axis): **2D** – monolayer culture at passage 5; **3D** – fibrin cultures; **Native** – whole tendon tissue. Boxplots present the distribution of the linear transformed C_t data (2^-dCt, y-axis). Tenomodulin (*Tnmd*) expression was significantly lower in fibrin than in native tendon. Although not represented in the differential expression analysis scleraxis, *Scx,* a key regulator of tenogenesis, was also considered. Scleraxis, and Mohawk (*Mkx*) expression was not significantly different across conditions, however, expression was low in fibrin cultures.
Significance code: * - $p<0.05$; ** - $p<0.01$; *** - $p<0.001$.

2.3.8: Cell viability in alginate bead cultures

A trypan-blue exclusion assay was used to define dead chondrocytes or tenocytes in alginate beads over a period of fourteen days, **Figure 2.12, SD2.27**. There was a statistically significant increase in the number of chondrocytes showing positive staining at all time points relative to the zero time point ($p<0.01$). There was no significant difference between chondrocyte and tenocyte values at any time point. Although the same trend for reduced viability with time was evident for tenocytes in alginate beads there was no significant reduction relative to the zero time point.

2.3.9: Profiling of homeobox genes in adult cartilage and tendon shows evidence of preserved anatomical topographical expression

The finding of differential expression of homeobox genes, including the hind-limb development associated *Pitx1*, prompted an investigation into the expression of homeobox genes associated with topographical anatomy in adult cartilage and tendon. The expression profiles of nineteen homeodomain genes, at four anatomical locations, were determined for cartilage and tendon. For the majority of genes a significant difference in gene expression between the four tissue sources could not be demonstrated, however statistically significant different expression was found for *Pitx1*, *Tbx4*, *Tbx5*, *Lmx1b*, *Tbx15*, *Hoxa13* and *Prrx2*, as presented in **Figure 2.13-2.15**.

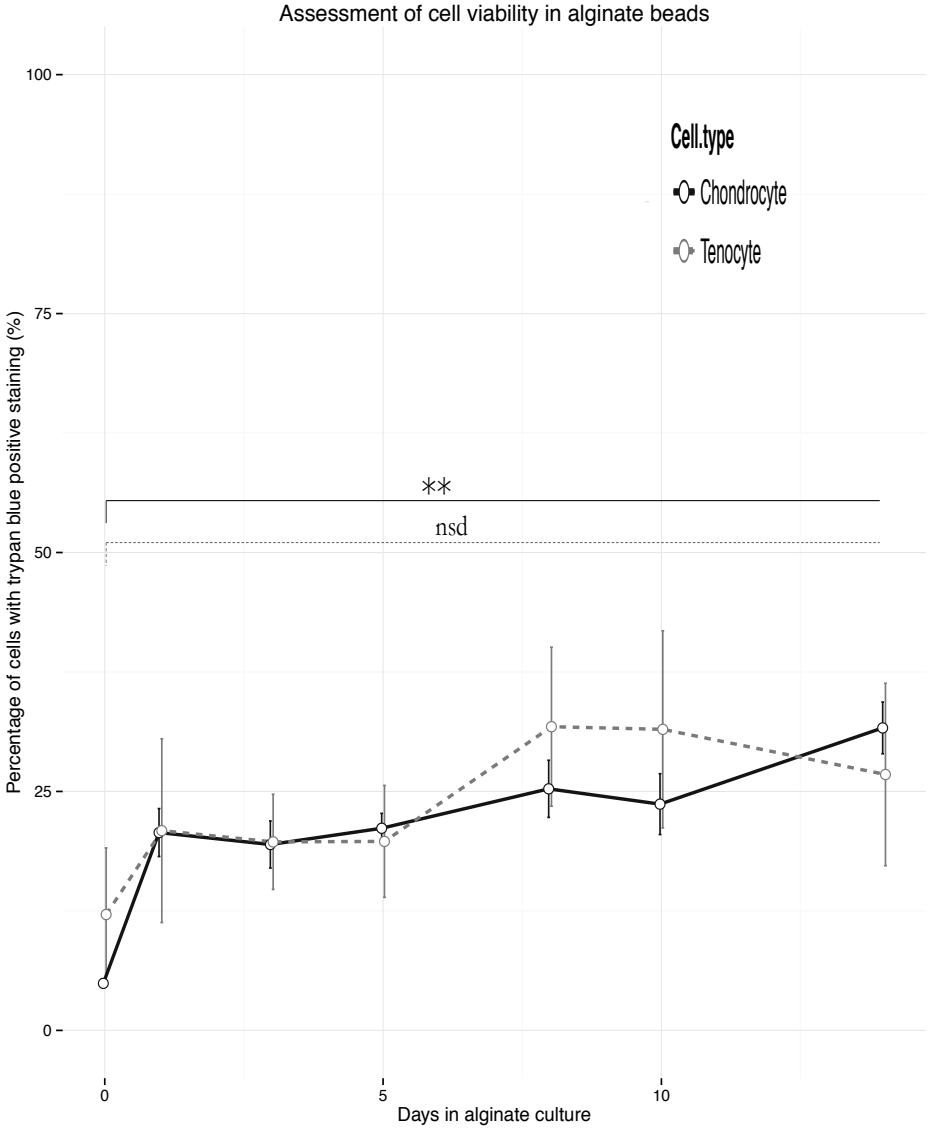


Figure 2.12: Viability assay for cells in alginate beads over a period of 14 days (x-axis) using a trypan-blue exclusion test. Data points show mean (n=4, technical triplicates) and standard error for percentage of positive/dead cells (y-axis) for chondrocytes (**solid**) and tenoytes (**dash**). Significance code – as before. A significant reduction in cell viability for chondrocytes in alginate beads was found between day 0 and all other time-points. Significance code: * - $p<0.05$; ** - $p<0.01$; *** - $p<0.001$; nsd – no significant difference.

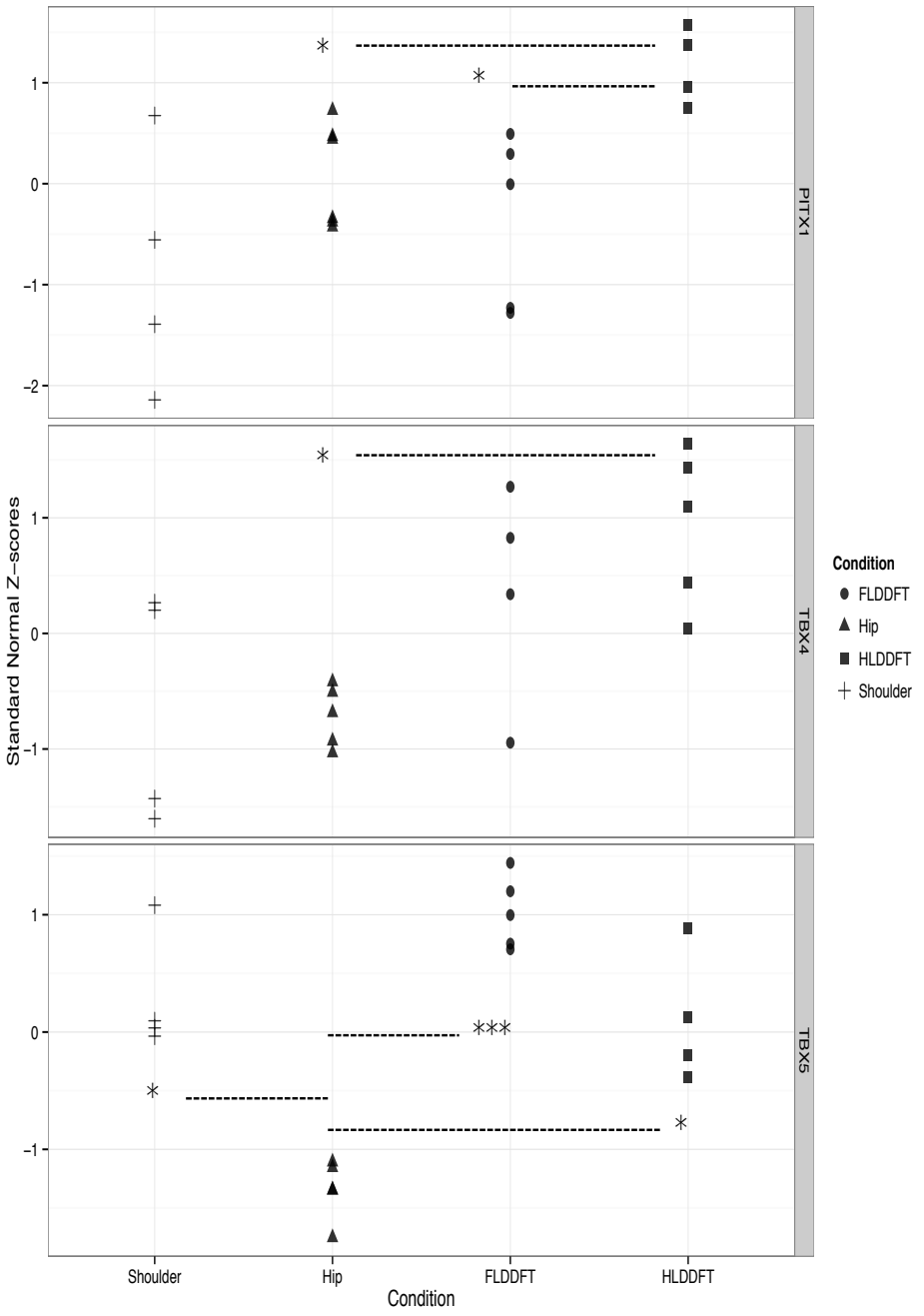


Figure 2.13: Z-scores are derived from the studentised, $\log_2$ transformed dCt values, higher *z*-scores are associated with higher expression. Hip and shoulder cartilage, along with forelimb (**FL**) and hindlimb (**HL**) deep flexor tendon (**DDFT**) are surveyed for differential expression of homeobox genes. Each biological replicate is represented in data points. Significance code: '*' – $p<0.05$; "**' – $p<0.01$; '***" – $p<0.001$. There was evidence of a significant difference in *Pitx1* expression between the fore- and hind-limb tendons. Hind-limb tendon demonstrated significantly higher expression of *Tbx4* relative to hip cartilage. Relative to hip cartilage *Tbx5* was more highly expressed in all other samples.

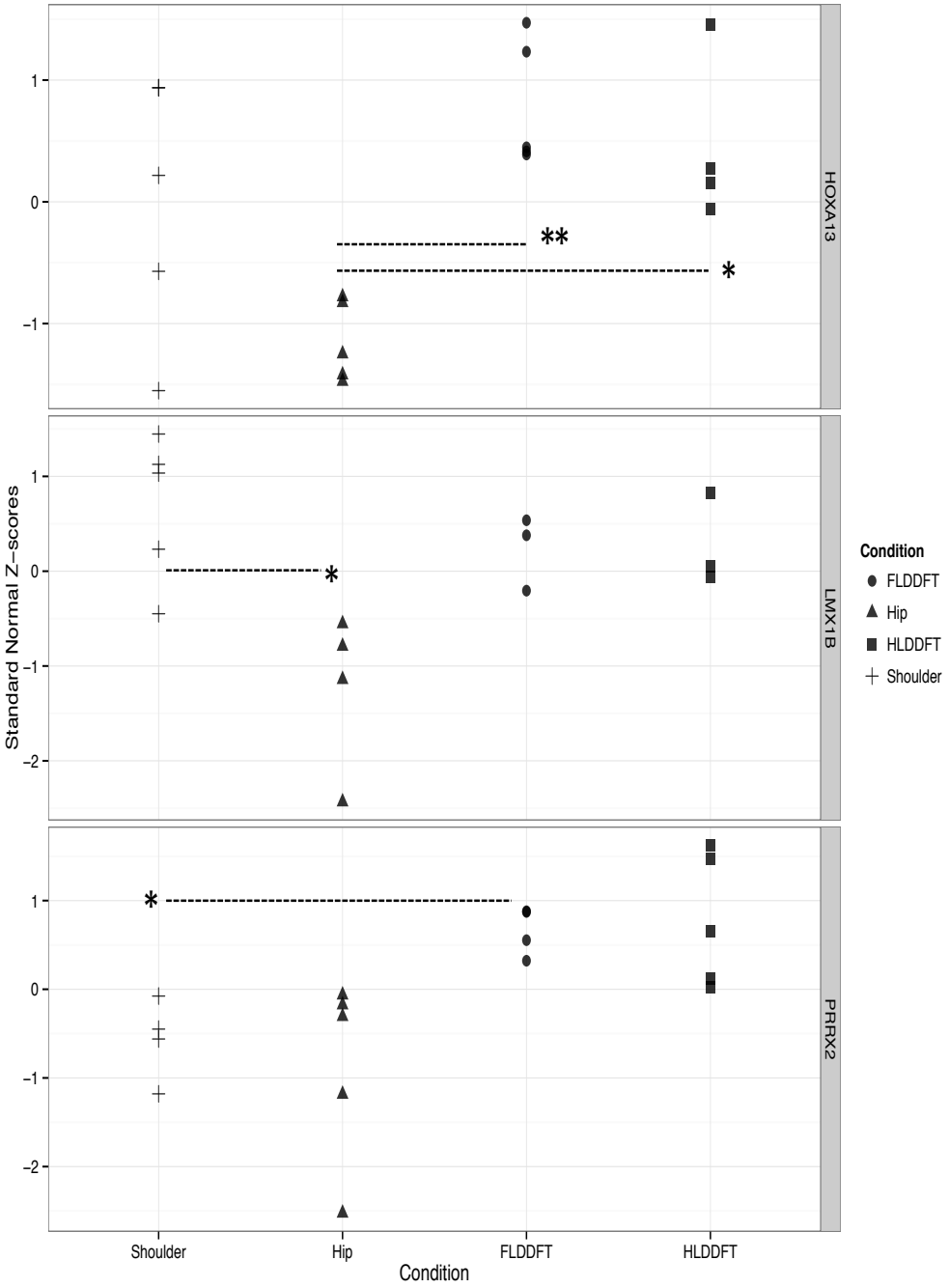


Figure 2.14: Z-scores are derived from the studentised, $\log_2$ transformed dCt values. Hip and shoulder cartilage, along with forelimb (**FL**) and hindlimb (**HL**) deep flexor tendon (**DDFT**) are surveyed for differential expression of homeobox genes. Significance code: '*' – $p<0.05$; "**' – $p<0.01$. Significant differences are found between hip cartilage expression of *Hoxa13* and tendon expression. Shoulder cartilage exhibits higher expression of *Lmx1b* than hip, but lower expression of *Prrx2* relative to the fore limb DDFT.

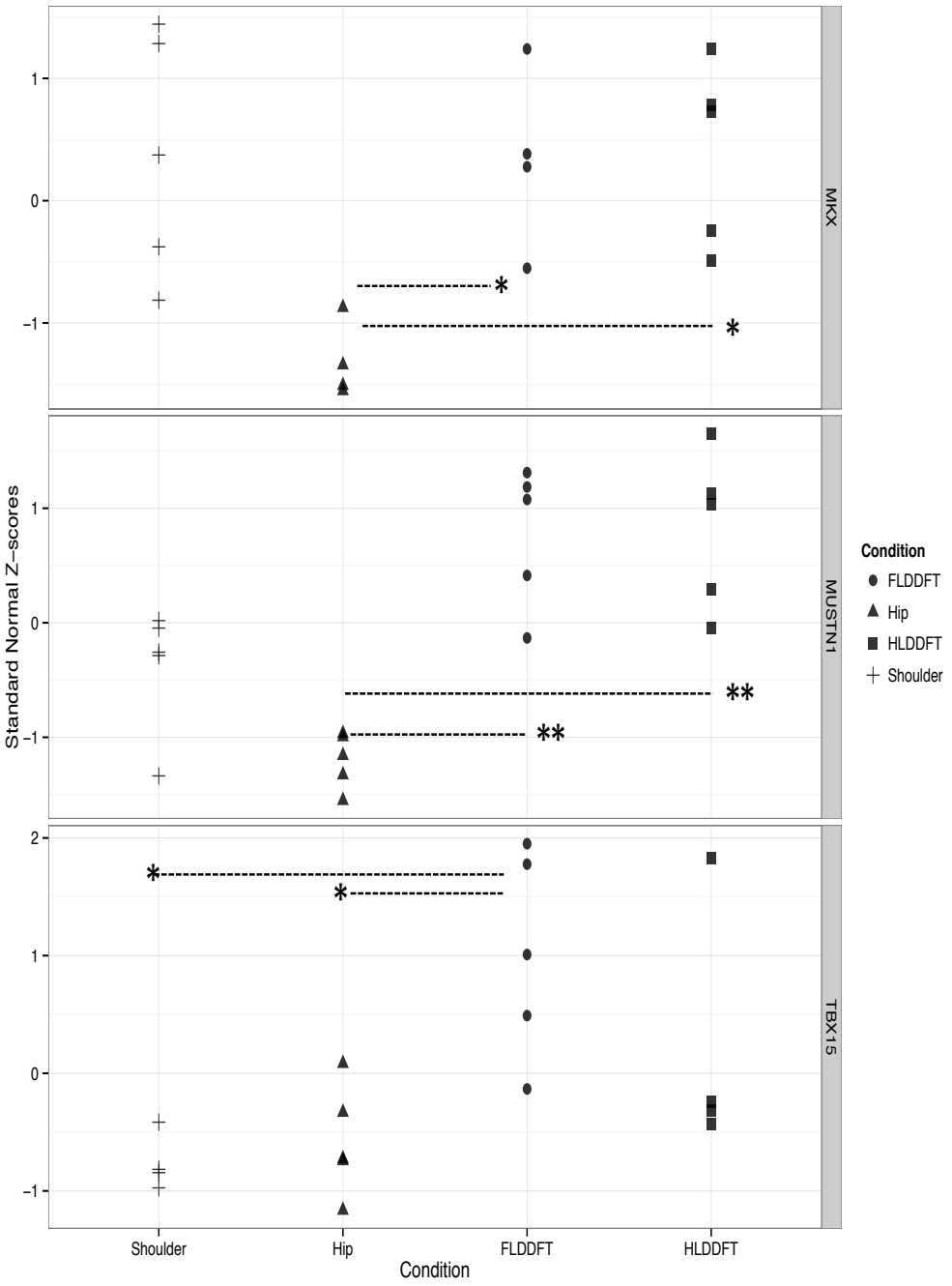


Figure 2.15: Z-scores are derived from the studentised, $\log_2$ transformed dCt values. Hip and shoulder cartilage, along with forelimb (**FL**) and hindlimb (**HL**) deep flexor tendon (**DDFT**) are surveyed for differential expression of homeobox genes. Significance code: '*' – $p<0.05$; "**' – $p<0.01$. Significant differences are found between hip cartilage and tendon for *Mustn1* and *Mkx* expression – both are more highly expressed in tendon. *Tbx15* is shown to be more highly expressed in fore limb DDFT than either cartilage source.

2.3.10: Histology and Immunohistochemistry

In order to validate findings from differential expression analysis in native tissue two target proteins were chosen. For tendon the fast skeletal muscle-associated troponin I, type 2 (encoded by *Tnni2*) was selected to define the presence of skeletal muscle-associated transcript expression in tendon. For cartilage the cathelicidin-associated anti-microbial peptide, *Cramp/Camp*, was considered as this represented a gene more commonly associated with polymorphonuclear cell expression and so would investigate the presence of innate immune responses in cartilage tissue or contamination from structures deep to the hyaline cartilage.

TENDON

It was not possible, using the described antibody and methodology, to categorically demonstrate specificity of staining for troponin I, fast skeletal muscle in tendon tissue, **Figures 2.16 to 2.18**. Strong, non-specific staining of the IgG isotype control (**Figure 2.17**) precluded further interpretation of low-level positive staining that was demonstrable in the myotendinous region of test studies (**Figure 2.18**). Furthermore, inconsistency of positive staining with tissue controls (associated skeletal muscle) did not allow for unambiguous interpretation of these results. With these sections, however, the deep invaginations of skeletal muscle into the Achilles tendon at the myotendinous junction were noted (**Figure 2.16**).

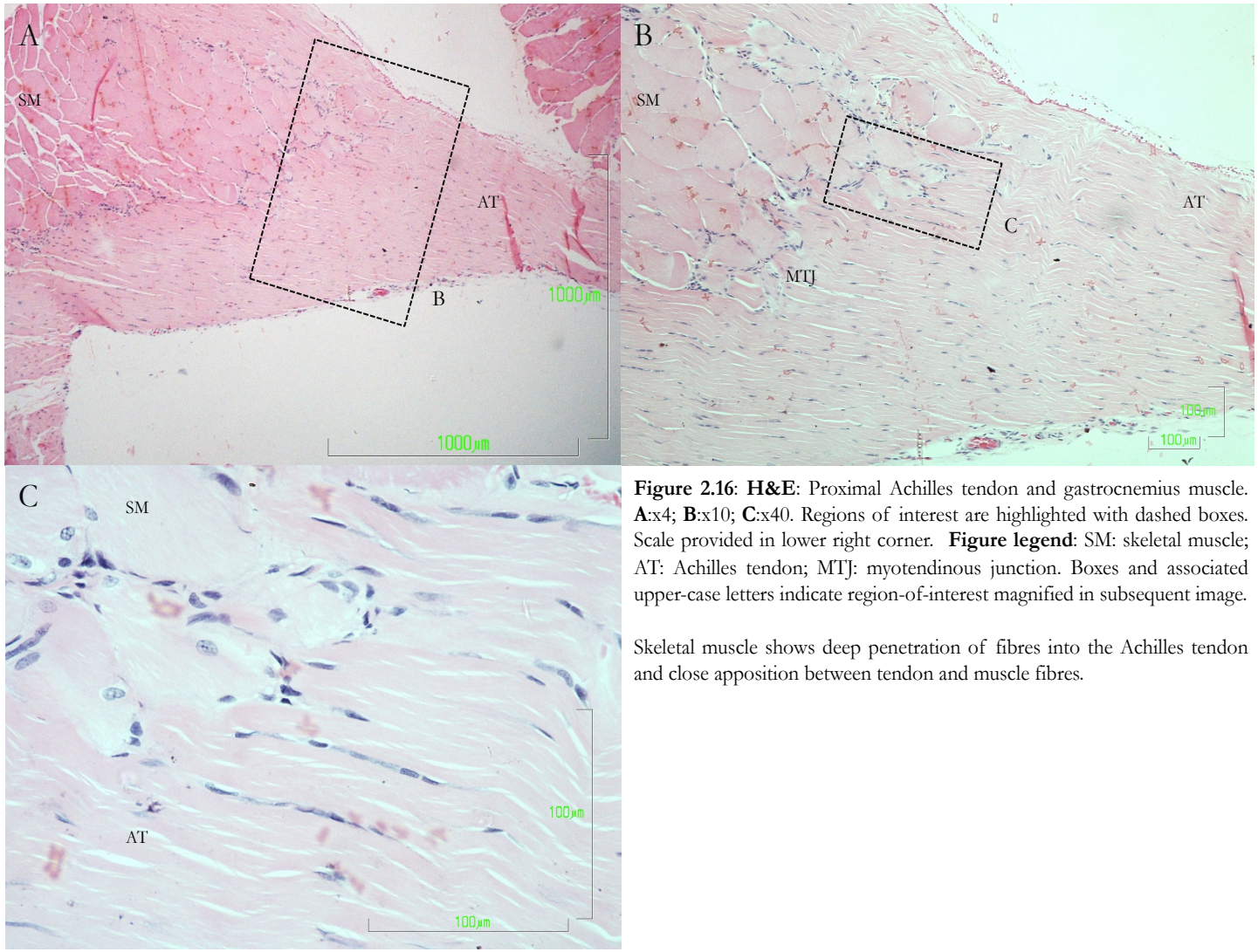


Figure 2.16: **H&E**: Proximal Achilles tendon and gastrocnemius muscle. **A**:x4; **B**:x10; **C**:x40. Regions of interest are highlighted with dashed boxes. Scale provided in lower right corner. **Figure legend**: SM: skeletal muscle; AT: Achilles tendon; MTJ: myotendinous junction. Boxes and associated upper-case letters indicate region-of-interest magnified in subsequent image.

Skeletal muscle shows deep penetration of fibres into the Achilles tendon and close apposition between tendon and muscle fibres.

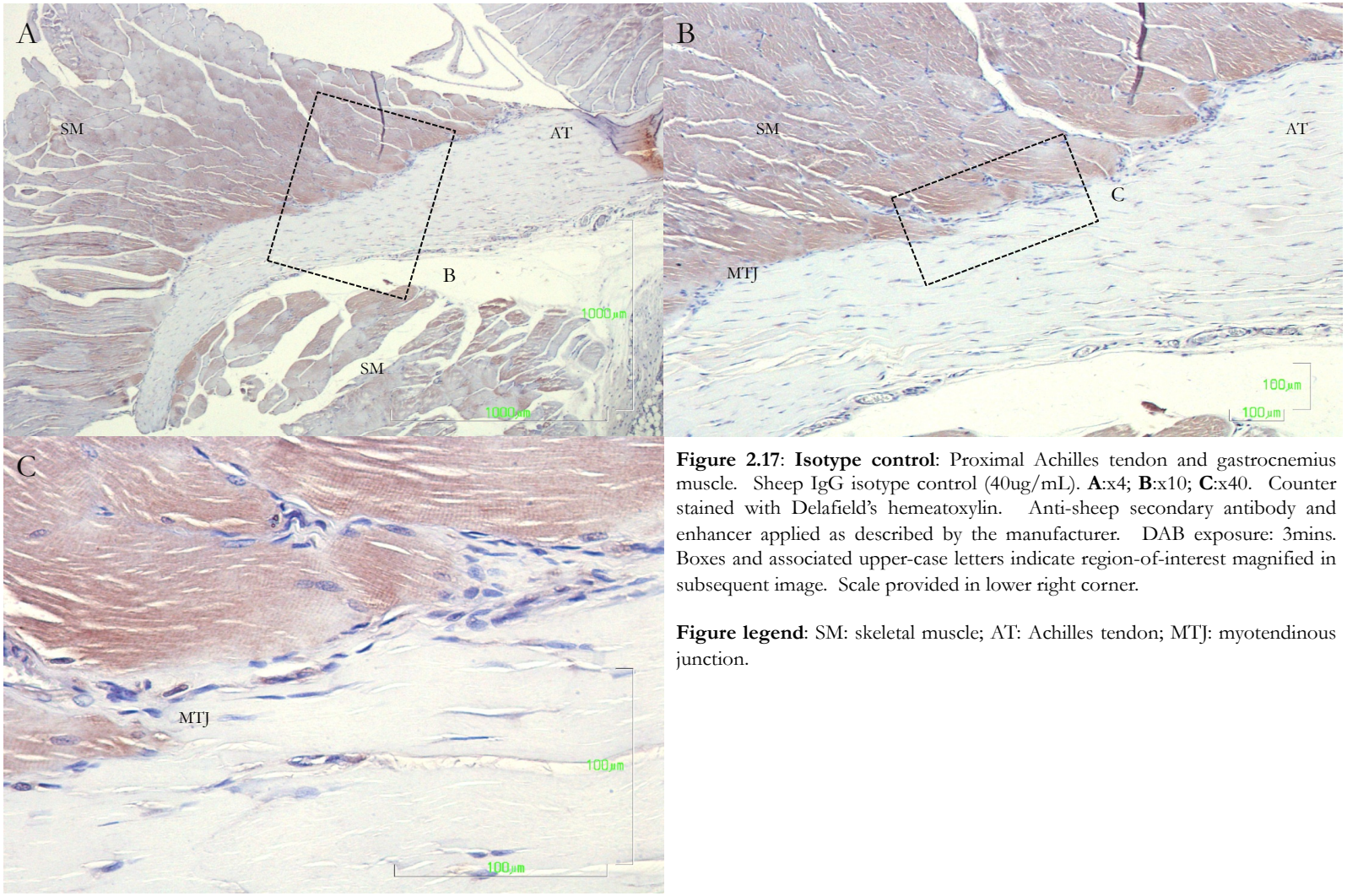


Figure 2.17: **Isotype control**: Proximal Achilles tendon and gastrocnemius muscle. Sheep IgG isotype control (40ug/mL). **A**:x4; **B**:x10; **C**:x40. Counter stained with Delafield's hemeatoxylin. Anti-sheep secondary antibody and enhancer applied as described by the manufacturer. DAB exposure: 3mins. Boxes and associated upper-case letters indicate region-of-interest magnified in subsequent image. Scale provided in lower right corner.

Figure legend: SM: skeletal muscle; AT: Achilles tendon; MTJ: myotendinous junction.

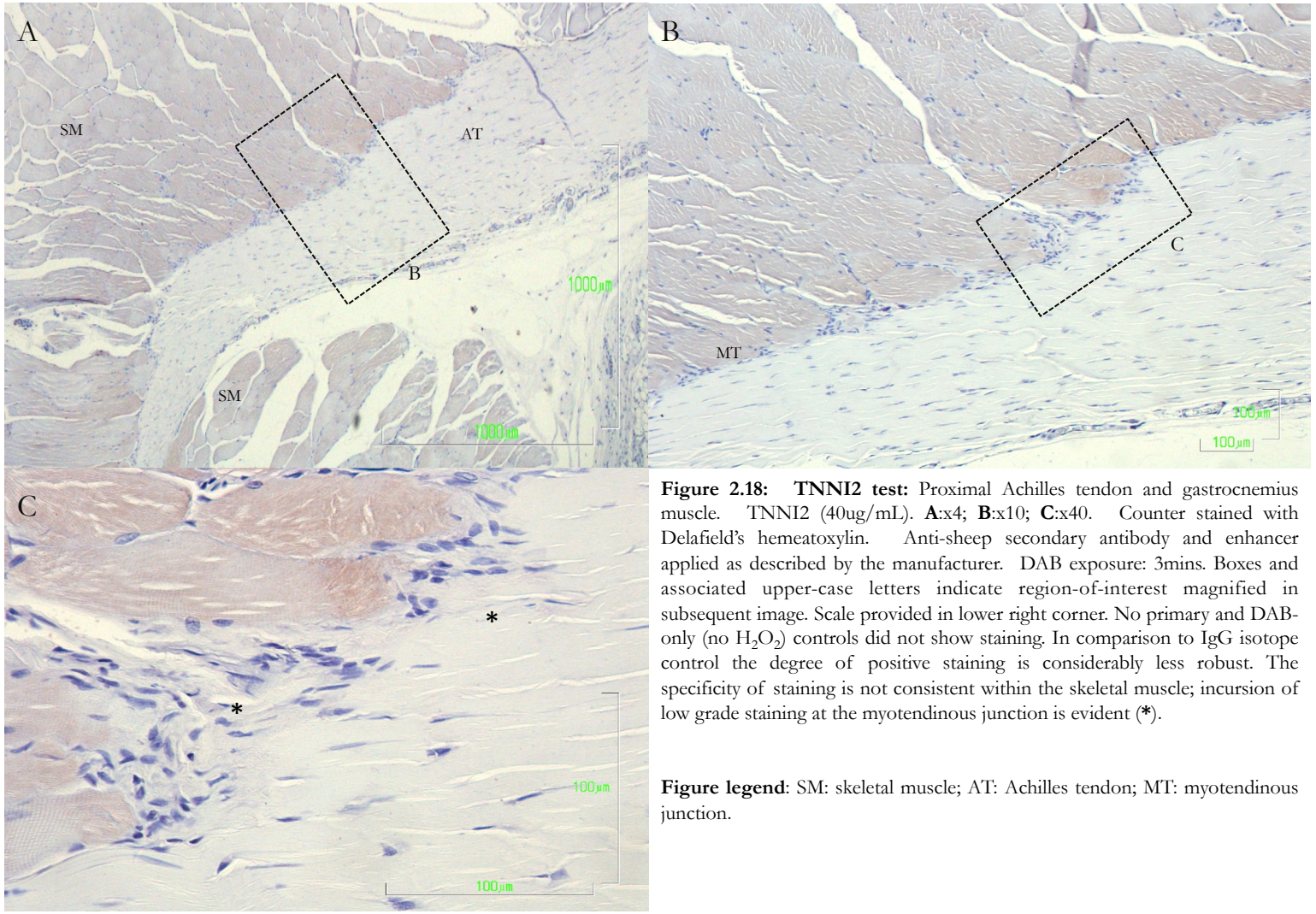


Figure 2.18: TNNI2 test: Proximal Achilles tendon and gastrocnemius muscle. TNNI2 (40ug/mL). **A**:x4; **B**:x10; **C**:x40. Counter stained with Delafield's hemeatoxylin. Anti-sheep secondary antibody and enhancer applied as described by the manufacturer. DAB exposure: 3mins. Boxes and associated upper-case letters indicate region-of-interest magnified in subsequent image. Scale provided in lower right corner. No primary and DAB-only (no H_2O_2) controls did not show staining. In comparison to IgG isotope control the degree of positive staining is considerably less robust. The specificity of staining is not consistent within the skeletal muscle; incursion of low grade staining at the myotendinous junction is evident (*).

Figure legend: SM: skeletal muscle; AT: Achilles tendon; MT: myotendinous junction.

Longitudinal sections of the rat femoro-tibial joint demonstrated the close association of sub-chondral blood vessels and the articular cartilage in the rat, **Figures 2.19-2.20**. In the proximal tibia there was evidence for the breaching of the osteochondral junction by blood vessels, which may infiltrate the calcified cartilage layer. As with the tendon samples high background staining with the isotype control (**Figure 2.21**) made interpretation of low-grade staining of the superficial layer chondrocytes with the CRAMP antibody difficult, **Figures 2.22**.

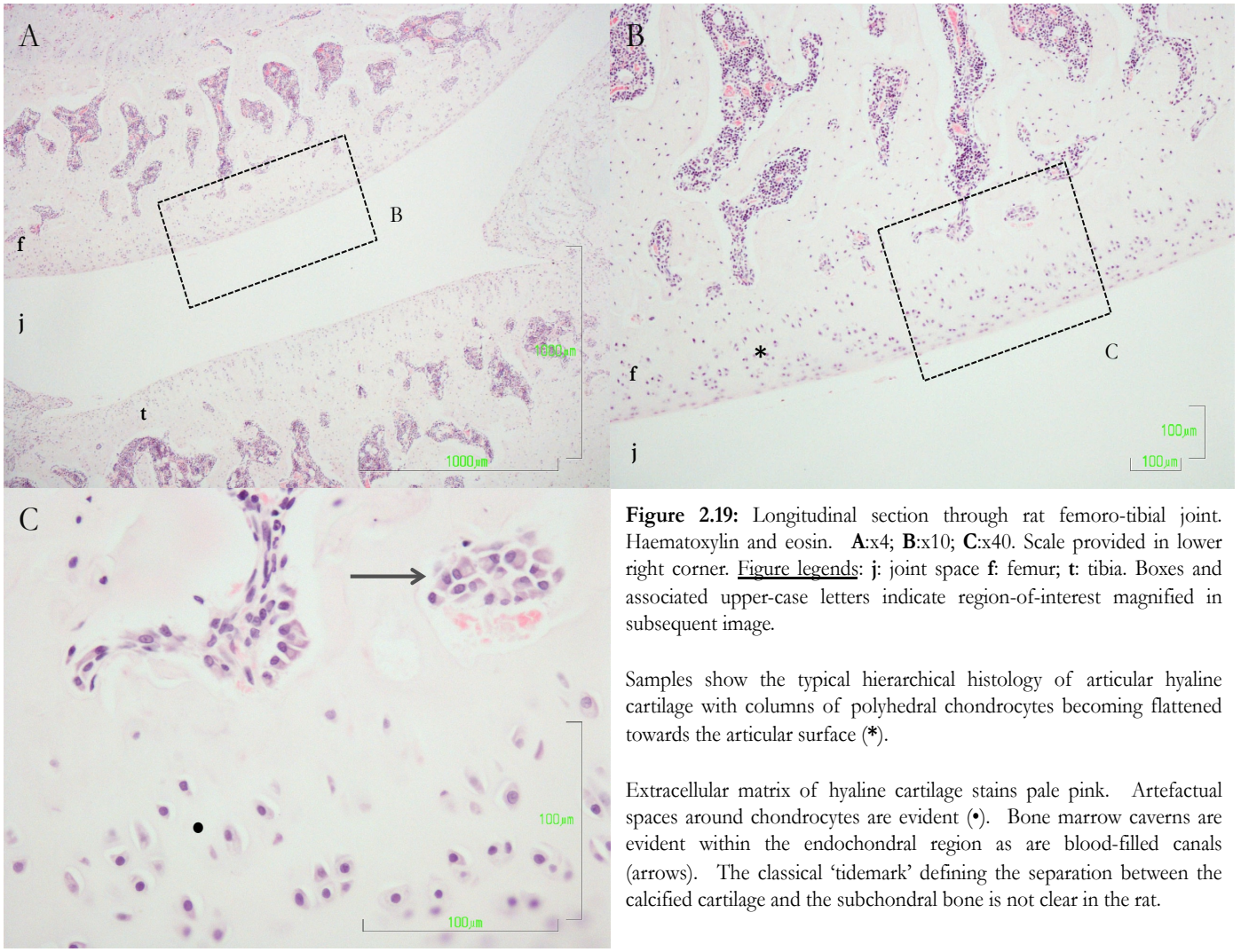


Figure 2.19: Longitudinal section through rat femoro-tibial joint. Haematoxylin and eosin. **A**:x4; **B**:x10; **C**:x40. Scale provided in lower right corner. Figure legends: **j**: joint space **f**: femur; **t**: tibia. Boxes and associated upper-case letters indicate region-of-interest magnified in subsequent image.

Samples show the typical hierarchical histology of articular hyaline cartilage with columns of polyhedral chondrocytes becoming flattened towards the articular surface (*****).

Extracellular matrix of hyaline cartilage stains pale pink. Artefactual spaces around chondrocytes are evident (•). Bone marrow caverns are evident within the endochondral region as are blood-filled canals (arrows). The classical 'tidemark' defining the separation between the calcified cartilage and the subchondral bone is not clear in the rat.

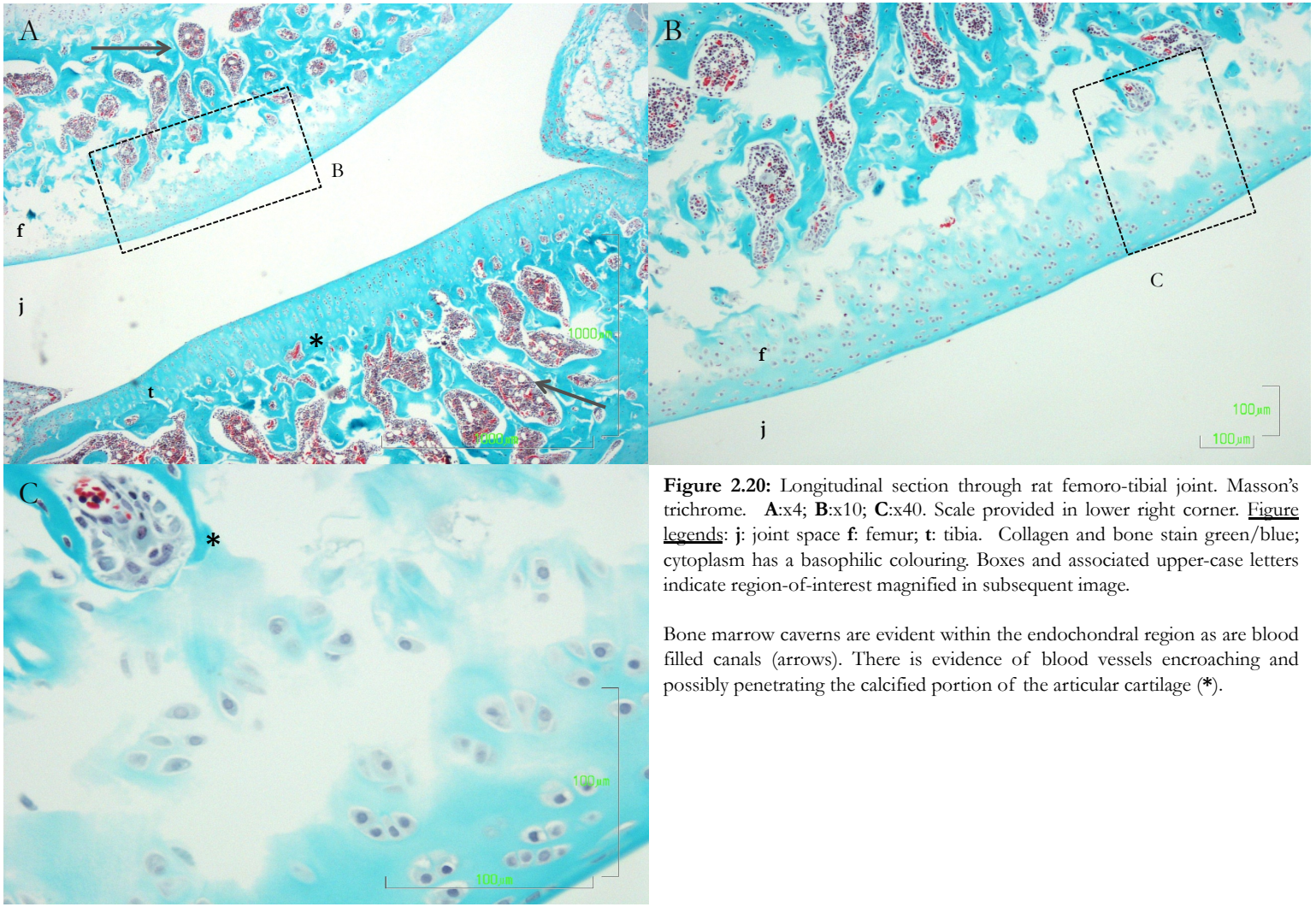


Figure 2.20: Longitudinal section through rat femoro-tibial joint. Masson's trichrome. **A**:x4; **B**:x10; **C**:x40. Scale provided in lower right corner. Figure legends: **j**: joint space **f**: femur; **t**: tibia. Collagen and bone stain green/blue; cytoplasm has a basophilic colouring. Boxes and associated upper-case letters indicate region-of-interest magnified in subsequent image.

Bone marrow caverns are evident within the endochondral region as are blood filled canals (arrows). There is evidence of blood vessels encroaching and possibly penetrating the calcified portion of the articular cartilage (*****).

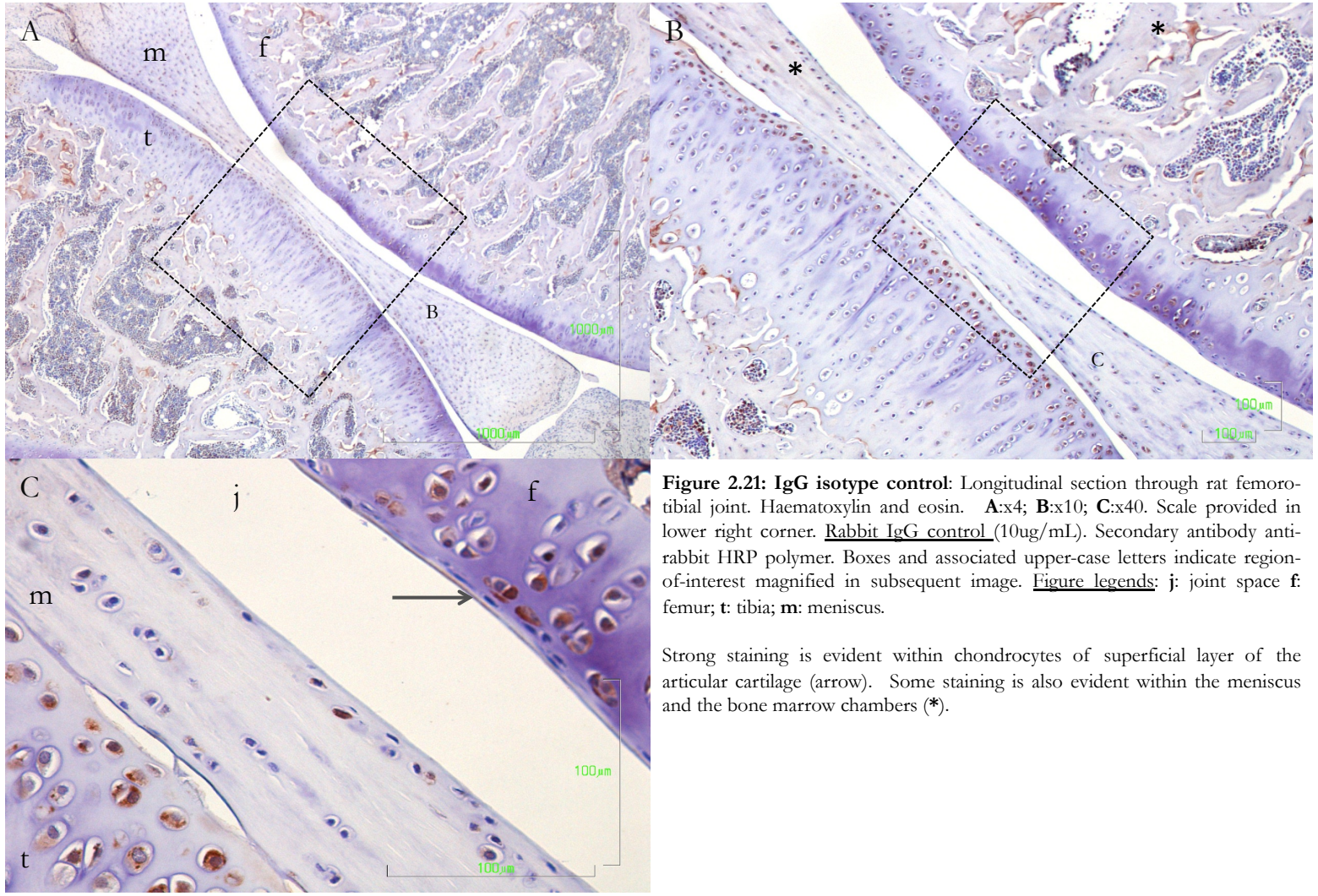


Figure 2.21: IgG isotype control: Longitudinal section through rat femoro-tibial joint. Haematoxylin and eosin. **A**:x4; **B**:x10; **C**:x40. Scale provided in lower right corner. Rabbit IgG control (10ug/mL). Secondary antibody anti-rabbit HRP polymer. Boxes and associated upper-case letters indicate region-of-interest magnified in subsequent image. Figure legends: **j**: joint space **f**: femur; **t**: tibia; **m**: meniscus.

Strong staining is evident within chondrocytes of superficial layer of the articular cartilage (arrow). Some staining is also evident within the meniscus and the bone marrow chambers (*).

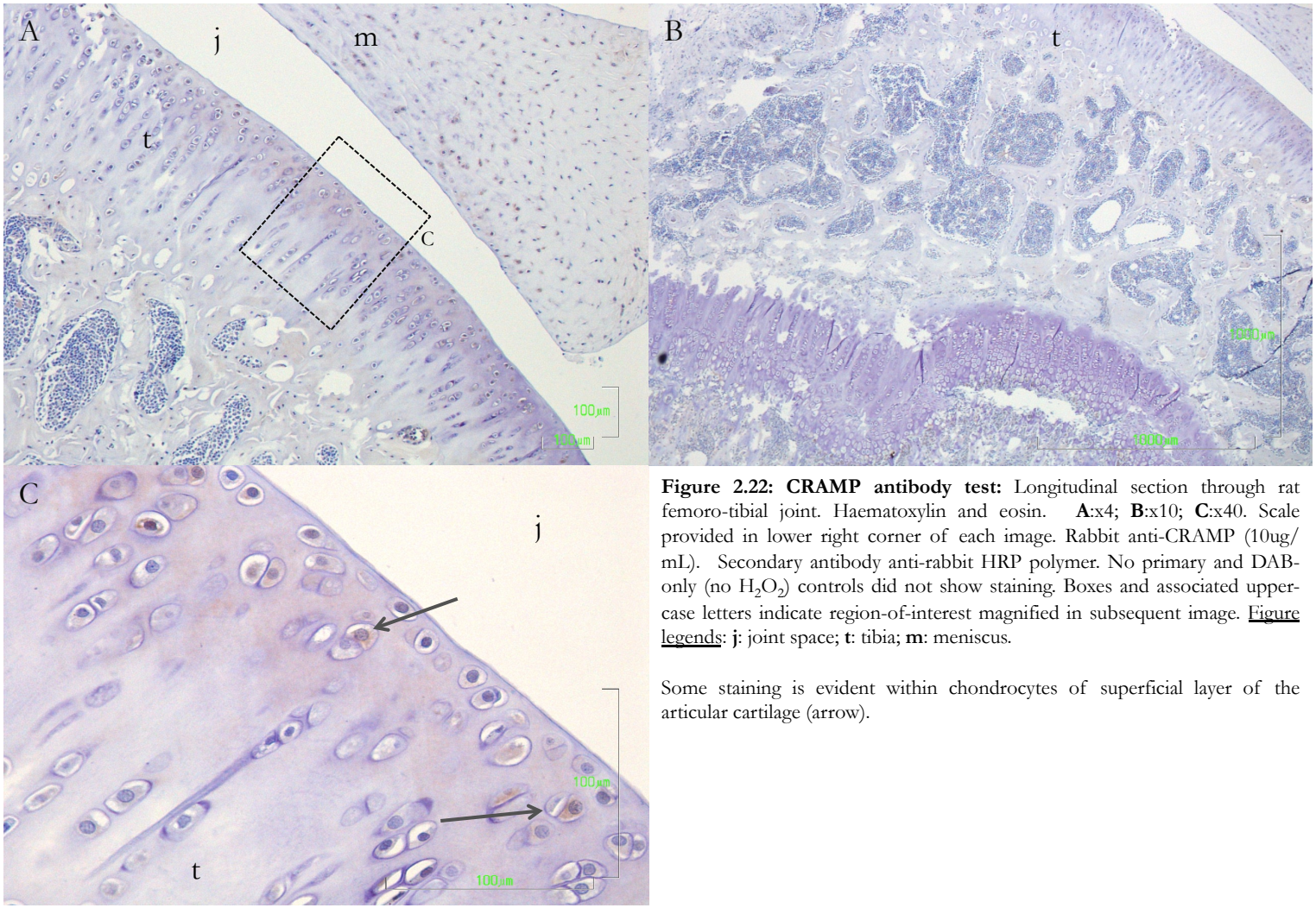


Figure 2.22: CRAMP antibody test: Longitudinal section through rat femoro-tibial joint. Haematoxylin and eosin. **A**:x4; **B**:x10; **C**:x40. Scale provided in lower right corner of each image. Rabbit anti-CRAMP (10ug/ mL). Secondary antibody anti-rabbit HRP polymer. No primary and DAB-only (no H_2O_2) controls did not show staining. Boxes and associated upper-case letters indicate region-of-interest magnified in subsequent image. Figure legends: **j**: joint space; **t**: tibia; **m**: meniscus.

Some staining is evident within chondrocytes of superficial layer of the articular cartilage (arrow).

2.4: Discussion

This study explored the phenotypic plasticity of chondrocytes and tenocytes, through induction of dedifferentiation in monolayer and subsequent redifferentiation in standard three-dimensional cultures, by global gene expression analysis. By defining a reference data set of prioritized genes, implicated signalling pathways and infered upstream regulators across three conditions a clearer understanding of the mechanisms governing dedifferentiation was sought. These objectives contributed to the wider goal of exploration of mechanisms that may contribute to degenerative phenotypes in cartilage and tendon and could inform the rational development of organotypic culture systems in tissue engineering.

The gene expression profiles of native, monolayer and standard three-dimensional culture systems provided the first comparative data set of its kind. Novel results demonstrated an inadequate restitution of native tissue expression profiles by commonly used three-dimensional culture models and offered an alternative description. In addition, convergence of gene expression profiles in monolayer culture, and the expression of development-associated genes in these cells, suggested dedifferentiation could represent a permissive phenotype worthy of further investigation for re-generation of musculoskeletal tissues. In particular, the expression of a hind-limb development-associated homeobox gene, *Pitx1*, in monolayer chondrocytes was validated indicating that further investigation of homeobox genes in dedifferentiation is warranted.

2.4.1: Monolayer cell culture gene expression profiles converge

Dedifferentiation of chondrocytes and tenocytes in monolayer was defined by the lower expression of *Col2a1* and *Tnmd* respectively, the latter confirmed by qPCR.

A striking finding of this study was the convergence of the gene expression profiles for chondrocytes, tenocytes and dermal fibroblasts at passage five. Fewer than three-hundred genes with a statistically significant differential expression were found between chondrocytes and tenocytes, and of the most highly differentially expressed only twenty-five had a $\log_2$ fold-change greater than two.

Given that monolayer culture is a fundamental research tool there is little scrutiny or comparison of gene expression profiles across cell types. By comparing tenocytes and chondrocytes in parallel it was possible to demonstrate this convergence. It is important to note that although chondrocytes and tenocytes were often harvested from the same animal, and from an isogenic line, pairing of tissue samples for microarray was not a feature of the initial experimental plan and convergence based upon replicate origin is unlikely.

Tissues are complex as they comprise three-dimensional hierarchical structure often with heterogenous cell populations, a vascular and neural supply, with caveats related to cartilage and tendon structure. Given the reduction in the complexity of the monolayer environment this finding is perhaps not surprising – once a proliferative phenotype is induced, genes concerned with DNA replication and metabolism predominate. Although this phenomenon has not been described before for cartilage and tendon cells others have reported comparable expression profile convergence (Sandberg and Ernberg 2005, Zaitseva, Vollenhoven et al. 2006, Halfon, Abramov et al. 2011). In a meta-analysis of human, mouse and rat

transcriptome profiles Prasad, *et al* (2013) demonstrated that, regardless of the tissue origin, cells in culture were convergent in their expression profiles and divergent from their tissues of origin in an order of magnitude comparable to the difference between tissue types (Prasad, Kumar et al. 2013).

If phenotypic drift and convergence of gene expression profiles in monolayer culture is a consistent feature of proliferation of chondrocytes and tenocytes then concerns regarding the validity and veracity of findings from monolayer studies should be considered. Furthermore, given that periods of expansion in monolayer cell culture are a pre-requisite for autologous cell therapeutic interventions (Brittberg, Lindahl et al. 1994) in musculoskeletal disease the nature of the cells re-implanted may be, crucially, functionally divergent from those initially harvested.

There are few studies that explore the global gene expression profiles of chondrocytes or tenocytes in monolayer. Ma, *et al* (2013) reported 93 genes that were consistently expressed at lower levels in passage 2 and passage 8 cultured chondrocytes compared to passage 0 (Ma, Leijten et al. 2013). There was broad overlap with the findings presented here for native chondrocytes with *Chi3l2*, *Col2a1*, *Frzb*, *Sox9* and *Mmp3* showing reduced expression in monolayer; fewer were more highly expressed in monolayer and these included *Twist1*, *Gpnmb*, *Smad3*, *Pparγ*, and *Tagln*. Long-term culture (up to three weeks) of tenocytes in monolayer was found to result in a significant reduction in *Tnmd* expression across all time points, but variable expression of *Scx* (Güngörmüş and Kolankaya 2012), consistent with the qPCR findings presented in this study.

Failure to define statistically significant differences in gene expression in all gene chosen for qPCR validation may have arisen from only using three biological

replicates per gene, use of ill-defined and heterogenous cell populations, primers of variable efficiency, or there indeed being no difference in expression. Clarification with larger, independent samples from enriched sub-populations may be required.

2.4.2: Establishing a qualitative definition for dedifferentiated cells

Given that chondroyctes and tenocytes in monolayer culture exhibit gene expression convergence some attempt must be made to define what this phenotype is; to concur with the definition of dedifferentiation as a regenerative mechanism this phenotype should be pre-differentiated and lineage-associated, i.e. represent a primitive musculoskeletal cell.

Comparisons with mesenchymal stem/stromal cells (MSC) was not a feature of this study, however the increased expression of a number genes associated with mesenchymal condensation and MSC identification indicated that the reduction in the expression of markers of differentiated status was not just a consequence of the reduction in the complexity of the samples, rather there was a global divergence from the native chondrocyte/tenocyte phenotype towards a homogenous, dedifferentiated phenotype.

The expression of the MSC cell-surface marker *Thy-1*/CD90 (Dominici, Le Blanc et al. 2006, Maleki, Ghanbarvand et al. 2014) and *Prnp*/prion gene expression, generally associated with embryonic stem cell differentiation studies (Lee and Baskakov 2010, Miranda, Pericuesta et al. 2011), in chondrocytes prompted consideration of dedifferentiation in monolayer as representing a proliferative 'pre-differentiated' state in the chondrocyte. For this to be the case it may be expected that overlap between developmental musculoskeletal gene expression profiles and

monolayer expanded chondrocytes and tenocytes might occur. Although a crude approach, the differential expression list for cartilage *versus* monolayer chondrocytes presented in this study shared 159 genes with the study by (Cameron, Belluoccio et al. 2009) considering the transcriptomic profile of *in vivo* murine chondrogenesis between E11.5 and 13.5. These genes included: *Pitx1*, *Thy1*, *Wisp1*, *Six1*, *Scrg1*, *S100a*, *Myh10*, *Il16*, *Frzb* and *Ctgf,* which have been presented in this study.

Applying this 'paint-by-numbers' approach to results from the first published global gene expression study of tendon development by Havis, *et al* (2014) reveals a number of genes differentially expressed between E11.5 and 14.5 familiar to the differential analysis for tendon *versus* monolayer tenocytes (Havis, Bonnin et al. 2014). A full gene list was not available, however, *Tnmd*, *Dpt*, *Clu, Mfap5, Thy1,* and *Bgn,* which were associated with higher expression at E14.5, were also differentially expressed in this study. Notably *Scx* was not a differentially expressed gene in the Havis study, but differential expression of cartilage-associated *Ibsp*, *Ogn* and *Postn* in tendon development was found.

These are general observations and do not attempt to equate monolayer chondrocytes with developing chondrocytes, or tenocytes, *per se*, rather suggest it may indicate a common gene regulatory network that facilitates a proliferative and permissive phenotype with potential for regenerative interventions.

Given the relevance of the development paradigm to musculoskeletal disease and regeneration (Tchetina 2011, Connizzo, Yannascoli et al. 2013) further consideration should be given to the regulatory networks that govern them. Clearly, to develop this would require comparisons between culture systems,

mesenchymal stem cells, and developing and healing tissues at the transcriptomic and proteomic level.

2.4.3 Gene expression characteristics of standard 3D cultures

Two forms of three-dimensional culture systems, alginate beads (De Ceuninck, Lesur et al. 2004) for chondrocytes, and fibrin cultures (Kapacee, Richardson et al. 2008) for tenocytes, fail to restitute an organo-typic profile in dedifferentiated cells from monolayer cultures. As these cultures do not show parity with cartilage and tendon it is critical that their characteristics are defined to facilitate refinement in tissue engineering.

Alginate bead cultures: chondrogenic and inflammatory profiles

The iterative analysis of gene network perturbations is critical to the development of rationally devised organo-typic models for *in vitro* use (Birgersdotter, Sandberg et al. 2005) and for the modelling of gene regulatory networks. To this end a global gene expression analysis of chondrocytes in three-dimensional alginate cultures was performed as an environmental perturbation. Gene expression profiles of passaged chondrocytes suspended in alginate beads under normoxic conditions reflected both chondrogenic and inflammatory profiles. As alginate cultures are extensively dealt with elsewhere in this thesis they are not the focus of this discussion.

Alginate expression profiles are suggestive of pro-inflammatory conditions with the expression of *Cox2*, *Cxcl1*, *Il*-6 and the alarmin family of osteoarthritis-associated genes (*S100b*) all evident. Chitinase 3-like protein 1 gene (*Chi3l1*) was the most highly expressed gene in alginate beads when compared to monolayer and it is a potential biomarker for osteoarthritis (Huang and Wu 2009). Recently,

CXCL1 has been defined as a target of the SOX9-mediated transcription factor AP-2ε (Wenke, Niebler et al. 2011), which is known to have a role in late-stage chondrocyte differentiation. Elsewhere, the expression of chemokines and their receptors has been demonstrated to be involved in the chondrogenic induction of MSCs (Cristino, Piacentini et al. 2008).

The S100-family of small-molecular weight calcium binding proteins are strongly represented in alginate cultures (*S100a1*, *S100a4*, *S100a16*, *S100a11*, *S100b*), several of which are identified in normal articular cartilage, but also up-regulated and associated with osteoarthritis (Yammani 2012). S100b and S100a4, for example, both are considered to have catabolic effects on matrix through the increased production of MMP13. The extracellular functions of the S100 proteins are mediated through their interaction with RAGE (receptor for advanced glycation end products), or Toll-like receptors, activation of which can influence a number of signalling cascades including, MAP kinases (p38, ERK1/2), PI-3K and NF-κB. In contrast to the work of Diaz-Romero, *et al* (2014) (Diaz-Romero, Quintin et al. 2014), this analysis found that there was an increase in expression of S100 genes in monolayer and alginate cultures, except for *S100a8*, which was more highly expressed in native cartilage. No *S100* genes were differentially expressed in a pairwise comparison of monolayer and alginate cultures. In an elegant study comparing the effects of gain- or loss-of-function using combinations of *S100a1* and *S100b*, Saito, *et al* (2007) demonstrated that over-expression inhibited chondrogenic differentiation; in contrast loss-of-function, by siRNA transfection, enhanced terminal differentiation (Saito, Ikeda et al. 2007). The S100 proteins are implicated in cartilage homeostasis, associated with disease phenotypes, and conflicting reports exist as to the expression status

of these genes in chondrocytes in various culture conditions. In isolation gene expression profiles alone do not appear to be sufficiently consistent to warrant their labelling as differentiation markers.

Expression of *Bhlhb2/Dec*-1 (Shen et al., 2002), *Gdf15/Mic*-1 (Iliopoulos, Malizos, Oikonomou, & Tsezou, 2008), *Srg1* (Ochi, Derfoul et al. 2006) and *Id1* (Asp, Thornemo, Inerot, & Lindahl, 1998) in alginate-encapsulated chondrocytes are also of interest as these all have implicated roles in cartilage development and/or disease. Along with the implicated signalling pathways there are aspects of both development- and disease-associated profiles present in the expression profile of chondrocytes in alginate beads. A number of homeobox genes, with chondrogenic associations (*Prrx1*, *Prrx2*, *Pitx1* and *Six1*), were more highly expressed in alginate bead cultures suggestive of a chondrogenic phenotype.

Global gene expression studies of alginate bead cultures have not been undertaken. These indicate a profile that is both pro-chondrogenic and pro-inflammatory and consists of known osteoarthritis-associated genes. Further work is required to elucidate the mechanism underlying this phenotype.

Fibrin cultures are associated with micro-fibril and tendon developmental gene expression markers

Previously fibrin constructs have been shown to demonstrate the presence of embryonic fibripositor structures (Kapacee, Yeung et al. 2010). Whilst this type of construct does not recapitulate the complexity of tendon it may serve as a valuable model of tendon development. Gene expression findings supporting this statement are explored below.

From the outset it was stipulated that dedifferentiation required evidence of a loss of terminal markers of differentiation; the significant reduction in the expression of *Tnmd* supports this statement.

Tnmd is considered a late marker of tendon cell differentiation (Liu, Zhu et al. 2014) and, as such, may explain the lack of expression in fibrin constructs. Instead more subtle markers of tenogenic induction should be considered and comparison to chondrocytes in alginate culture aids this investigation. Relative to alginate cultures cells in fibrin constructs expressed higher levels of *Slit3*, *Thbs4*, *Mfap5*/*Magp2*, *Meox2* and *Tnn*, which may represent tenogenesis markers for further validation. Three of these novel findings in fibrin cultures are discussed further.

Microfibril-associated gene expression is higher in fibrin cultures

Microfibril-associated glycoprotein 2 (*Magp2*) is part of the fibrillin-based microfibril complexes, which have key roles in tissue integrity and elastic structure (Gibson, Finnis et al. 1998). In this study higher expression of *Magp2*/*Mfap5* is shown in fibrin constructs and native tendon relative to monolayer tenocytes and alginate cultures, but not relative to each other. Expression is present in a number of tissues, but high levels of mRNA have been noted in the foetal bovine Achilles tendon (Gibson, Finnis et al. 1998). Originally shown to be co-expressed in the foetal nuchal ligament, *Magp2* and *Magp1* are associated with fibrillin-containing microfibrils in the ECM of a number of tissues, which are variably associated with elastin fibres. Ritty, *et al* (2002) considered the distribution of *Magp2* in the deep flexor tendon and found that both microfibril-associated glycoproteins were distributed throughout the tendon, particularly at the insertional zones (Ritty, Ditsios et al. 2002). Roles for *Magp2* are less well described than *Magp1*. The

availability and activity of some growth factors (TGF-β, BMP, Notch family members) are regulated by microfibril assemblies, which provide a specific context for TGF-β and BMP signaling (Ramirez and Rifkin 2009). Fibrillins, mutations of which result in Marfan syndrome, have a role in the sequestration of growth factors in large latent complexes within the ECM. *Magp1* has been shown to have an inhibitory effect on the binding of latent TGF-β-binding protein 1 to fibrillin-1 in a study of the incorporation of latent TGF-β into the large latent complex in the ECM (Massam-Wu, Chiu et al. 2010). Whether MAGP2, which is evolutionarily related, but diverges structurally, has comparable functions is not well described. A recent loss-of-function study has shown that MAGP2 has wide-ranging effects not in keeping with the effects of MAGP1 or fibrillin-1 deficiency (Combs, Knutsen et al. 2013). Biochemical evidence suggests that MAGP2 protein binds to TGF-β1, TGF-β2 and BMP2. Unlike MAGP1 knock-out (k-o) mice, MAGP2 are not osteopenic, however double k-o mice do show enlarged aortic diameter suggesting that some combinatorial effect is in play to maintain the integrity of large, elastic blood vessels. With the recent finding of increasing levels of *Mfap5*/*Magp2* in E14.5 developing tendon (Havis, Bonnin et al. 2014) the higher expression in fibrin culture here may indicate a tenogenic profile.

Fibrin constructs show higher expression of tendon development-associated tenascin

Consistently in pairwise comparisons *Tnn*, encoding the tenascin-W isoform, was more highly expressed in fibrin constructs relative to native tendon, monolayer or alginate cultures. Tenascins are high molecular weight glycoproteins within the ECM, comprising of variable repeats of identical subunits (Halper and Kjaer 2014) with roles in cell motility, proliferation and differentiation. Their tissue distribution is variable but tenascin-C and –W show expression in a number of

developing structures and neoplastic stromal tissue (Tucker and Chiquet-Ehrismann 2009). Expression of tenascin-C in tendon blastema is well-described (Schweitzer, Chyung et al. 2001), but tenascin-W less so. In an avian study of the distribution of tenascin-W in development the predominant expression was localized in bone and perisoteum, but was also transiently expressed in smooth muscle, tendon and ligament, often overlapping with tenascin-C expression (Meloty-Kapella, Degen et al. 2006). The adhesion modulating properties of tenascin-W and lack of a connective tissue phenotype for the tenascin-C knock-out suggests that there may be some functional overlap. Tenascins are known to have adhesion-modulating properties (Chiquet-Ehrismann and Tucker 2011) and recently this has been confirmed for tenascin-W (Brellier, Martina et al. 2012). In this study, using a mouse C2C12 myoblast cell line with osteoblast differentiation potential, it was shown that cells cultured in the presence of tenascin-W maintained a stellate phenotype with pseudopodia, unlike the spreading and stress-fibres that formed when cultured on fibronectin alone.

Slit3, with a role in axonal guidance, shows higher expression in monolayer and fibrin cultures

In monolayer cultures a higher expression of *Slit3* is found relative to native tendon and in fibrin cultures relative to alginate beads. By qPCR a trend towards higher expression in fibrin constructs was presented. The SLIT protein family ligands, and their receptors, ROBO proteins, are required for normal axon guidance during development in vertebrates (Hinck 2004) and, additionally, myogenesis and myotendinous junction formation in *Drosophila* (Kramer, Kidd et al. 2001, Gilsohn and Volk 2010, Krämer, Green et al. 2014). *Slit3* deficiency in mice has been demonstrated to result in congenital diaphragmatic hernia as a result of failure of the central tendon, which helps suspend the diaphragm from

the body wall (Liu, Zhang et al. 2003). The pattern of *Slit* expression in the developing avian and murine limb would indicate that SLIT:ROBO have additional actions. In the developing chick *Slit3* was demonstrated to be strongly expressed in the dorsal and ventral central mesenchyme and prospective inter-digital regions (Holmes and Niswander 2001); later this expression developed along the nascent digit borders (Vargesson, Luria et al. 2001). Whilst *Slit1* and *Slit2* co-localised with the myoblast marker *Pax7* the subectodermal *Slit3* domains did not. Attempts to co-label tenascin-positive cells, to define developing tendon populations, could not localise a particular *Slit* population.

The relevance of vertebrate SLIT proteins to tendon development are derived from the role of the single gene ortholog, *slit*, in Drosophila, which is secreted by tendon progenitors to influence myotube migration through the ROBO receptor (Volk 1999, Gilsohn and Volk 2010, Schweitzer, Zelzer et al. 2010) under the influence of the EGR-like transcription factor homolog, *Stripe* (Volohonsky, Edenfeld et al. 2007). Although early-growth response proteins have been shown to be associated with tendon development (Lejard, Blais et al. 2011) in this data set higher expression of *Egr1* was found in alginate beads relative fibrin cultures. Whether the signaling mediators of axonal guidance and neuronal path-finding have a role in mammalian neo-tendon formation has yet to be elucidated, but there is a rational scope for further work in this area.

With respect to the developmental associations of the selected genes in the literature there is a rationale to explore further the validity of fibrin constructs as a tendon development model. As with alginate cultures, side-by-side gene expression studies with embryonic tendon and MSCs would provide valuable information on de- and re-differentiation mechanisms in tenocytes.

2.4.4: Pathway topology related to fibrin cultures

Predicted signalling pathways have not been described for tendon fibrin constructs. In this section one of the identified signalling pathways is discussed further with respect to its role in tendon physiology.

HIF1 signalling

Pathway topology predicts the activation of the HIF1-signalling pathway in the transition from monolayer to fibrin cultures, and in the theoretical comparison of fibrin cultures to native tendon. Expression of the *Hif-1α* gene, encoding the alpha subunit of HIF1, is more highly expressed in fibrin constructs, cultured in normoxic conditions, relative to either monolayer or native tendon. The response to changes in oxygen-tension during development and disease process is, partly, regulated by the DNA-binding HIF transcription factors. The transcriptional activation of multiple genes is mediated by the binding of HIF heterodimers, of which the alpha-subunit is the oxygen-labile component (Brocato, Chervona et al. 2014). The role of HIF-1α in human shoulder tendinopathy has been reported by several groups (Benson, McDonnell et al. 2010, Lakemeier, Reichelt et al. 2010), but studies are not always well controlled and immunohistochemistry findings were equivocal. Recently, Miller, *et al* (2012) presented higher mRNA and protein expression of HIF-1α in early subscapularis tendinopathy and passaged tenocytes cultured in hypoxic conditions (Millar, Reilly et al. 2012). The authors proposed that hypoxic cell injury was a critical element in the pathophysiology of tendinopathy. Higher expression of clusterin was also demonstrated.

Gene expression profiling of fibrin cultures demonstrates the higher expression of a number of pro-inflammatory cytokines (*Ptgs2/Cox2*, angiogenic factors (*Vegf-B,*

Angptl4, Angptl2), HIF-repressors (*Cited2*) and pro-apoptosis genes (*Bnip3*) some with known tendon pathology associations.

As such, it is proposed that rather than representing an organotypic *in vitro* model of healthy tendon structure the fibrin construct does not adequately recapitulated the native tendon expression profile; in contrast, it represents either a regenerative/healing, developmental or degenerate phenotype. This is consistent with the conclusions reached for the alginate bead model. Neither culture system appears fit for purpose with respect to native tissue, however, they may still be useful as either models of disease or development.

2.4.5: Histology and Immunohistochemistry

As a result of the presence of high levels of haemoglobin transcripts and various transcripts for genes more commonly associated with white blood cell function further consideration of the histological structure of the rat knee was warranted. The presence of high levels of *Hba* and *Hbb* are not uncommon in cartilage studies, although little is made of the presence in gene expression profiles.

Concerns relating to the infiltration of blood vessels into the articular cartilage are verified by Mapp, *et al* (2008) (Mapp, Avery et al. 2008) where breaching of the osteochondral junction by vascular channels in the tibial plateau is found to be a normal feature of the rat knee. Further, they found that the degree of vascularity in the tibial plateau decreased with age. Incursion of blood vessels into the non-calcified deep zone of articular cartilage appears to be a rare event in normal cartilage, however, perforation of the calcified cartilage does occur. Clark (1990) described the trajectory of sub-chondral vascular channels in rabbit, canine and human cartilage scanning electron microscopy (SEM) study (Clark 1990). These

SEM studies demonstrated that vascular channels, vertical extensions of 10-30 μm canals within the bone plate, variably penetrated the sub-chondral bone layer. These vascular channels either contained single capillaries with an open leading edge facing the calcified cartilage, or were, instead, covered by a cap of lamellar bone. Vascular channels coming into contact with the non-calcified articular cartilage was rare event. Contamination of native cartilage samples with blood cannot be excluded given the potential for encroachment of capillaries into the calcified cartilage in the rat.

Defensins and alarmins in normal cartilage

Key in this discussion is separating what is contamination of cartilage samples with blood and what is part of the normal physiological expression profile of whole cartilage. To this end the expression of defensins, antimicrobial peptides of the innate immune system, in whole cartilage must be considered. In this study the expression of the cathelicidin-related antimicrobial peptide (CRAMP/CAMP) was investigated by immunohistochemistry. Expression of this protein in normal chondrocytes and synoviocytes was reported by Hoffmann, *et al* (2013) and shown to be increased following pristane-induced arthritis in rats (Hoffmann, Bruns et al. 2013). In this study the same antibody was used, however, it was not possible to confirm the specific staining presented by Hoffmann, *et al* (2013). This study did not present control images, include isotype controls, nor define in the methods the antibody concentrations used. The high expression levels of alarmins, discussed above, and the expression of defensin genes requires further investigation in normal cartilage.

2.4.6: Evidence of topographical preservation of *Hox* gene expression indicates the heterogeneity of cartilage and tendon sources

Homeobox genes encode a coordinate system for limb patterning in development (Zakany and Duboule 2007). In a preliminary study, the first of its kind, the preservation of site-specific expression of a panel of homeobox genes was considered in cartilage and tendon. Aspects of the study design limited interpretation of the data including the confounding association between tissue and location - tendon (distal limb) and cartilage (proximal limb). The variable tissue harvest and RNA recovery and technical factors (multiple plates, temporal variation in data gathering, technical expertise) may also limit reproducibility. Despite these limitations a number of notable findings were evident, for example, the significantly higher expression of the hind limb-associated *Pitx1* (Marcil, Dumontier et al. 2003) gene in hindlimb tendon relative to the forelimb equivalent, or the higher expression of the forelimb associated *Tbx5* (Takeuchi, Koshiba-Takeuchi et al. 1999, Agarwal, Wylie et al. 2003) in shoulder cartilage relative to hip. Whilst these findings are encouraging further investigation would benefit from more targeted gene expression profiling of a variety of tissue sites (tail tendon, patellar tendon, axial cartilage) in a standardised manner.

References

Ackema, K. B. and J. Charite (2008). "Mesenchymal stem cells from different organs are characterized by distinct topographic Hox codes." Stem Cells Dev **17**(5): 979-991.

Agarwal, P., J. Wylie, J. Galceran, O. Arkhitko, C. Li, C. Deng, R. Grosschedl and B. Bruneau (2003). "Tbx5 is essential for forelimb bud initiation following patterning of the limb field in the mouse embryo." Development (Cambridge, England) **130**(3): 623-633.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nature genetics **25**(1): 25-29.

Ausma, J., M. Wijffels, G. van Eys, M. Koide, F. Ramaekers, M. Allessie and M. Borgers (1997). "Dedifferentiation of atrial cardiomyocytes as a result of chronic atrial fibrillation." The American Journal of Pathology **151**(4): 985-997.

Benson, R. T., S. M. McDonnell, H. J. Knowles, J. L. Rees, A. J. Carr and P. A. Hulley (2010). "Tendinopathy and tears of the rotator cuff are associated with hypoxia and apoptosis." The Journal of Bone and Joint Surgery. British volume **92**(3): 448-453.

Benya, P. and J. D. Shaffer (1982). "Dedifferentiated chondrocytes reexpress the differentiated collagen phenotype when cultured in agarose gels." Cell **30**(1): 215-224.

Bhattacharya, S., Q. Zhang and M. Andersen (2011). "A deterministic map of Waddington's epigenetic landscape for cell fate specification." BMC Systems Biology **5**(1): 85.

Birgersdotter, A., R. Sandberg and I. Ernberg (2005). "Gene expression perturbation in vitro—A growing case for three-dimensional (3D) culture systems." Seminars in Cancer Biology **15**(5): 405-412.

Bolstad, B. M., R. A. Irizarry, M. Åstrand and T. P. Speed (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-193.

Brellier, F., E. Martina, M. Chiquet, J. Ferralli, M. van der Heyden, G. Orend, J. Schittny, R. Chiquet-Ehrismann and R. Tucker (2012). "The adhesion modulating properties of tenascin-W." International Journal of Biological Sciences **8**(2): 187-194.

Brittberg, M., A. Lindahl, A. Nilsson, C. Ohlsson, O. Isaksson and L. Peterson (1994). "Treatment of Deep Cartilage Defects in the Knee with Autologous Chondrocyte Transplantation." N Engl J Med **331**(14): 889-895.

Brocato, J., Y. Chervona and M. Costa (2014). "Molecular responses to hypoxia-inducible factor 1 α and beyond." Molecular Pharmacology **85**(5): 651-657.

Cairns, J. M., M. J. Dunning, M. E. Ritchie, R. Russell and A. G. Lynch (2008). "BASH: a tool for managing BeadArray spatial artefacts." Bioinformatics **24**(24): 2921-2922.

Cameron, T., D. Belluoccio, P. Farlie, B. Brachvogel and J. Bateman (2009). "Global comparative transcriptome analysis of cartilage formation in vivo." BMC Developmental Biology **9**(1): 20.

Campbell, K. H. S., J. McWhir, W. A. Ritchie and I. Wilmut (1996). "Sheep cloned by nuclear transfer from a cultured cell line." Nature **380**(6569): 64-66.

Chang, H. Y. (2009). "Anatomic Demarcation of Cells: Genes to Patterns." Science **326**(5957): 1206-1207.

Chiquet-Ehrismann, R. and R. Tucker (2011). "Tenascins and the importance of adhesion modulation." Cold Spring Harbor: Perspectives in Biology **3**(5).

Chomczynski, P. and N. Sacchi (2006). "The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on." Nat. Protocols **1**(2): 581-585.

Clark, J. M. (1990). "The structure of vascular channels in the subchondral plate." Journal of Anatomy **171**: 105-115.

Combs, M., R. Knutsen, T. Broekelmann, H. Toennies, T. Brett, C. Miller, D. Kober, C. Craft, J. Atkinson, M. Shipley, B. Trask and R. Mecham (2013). "Microfibril-associated glycoprotein 2 (MAGP2) loss of function has pleiotropic effects in vivo." The Journal of Biological Chemistry **288**(40): 28869-28880.

Connizzo, B., S. Yannascoli and L. Soslowsky (2013). "Structure-function relationships of postnatal tendon development: a parallel to healing." Matrix Biology **32**(2): 106-116.

Cristino, S., A. Piacentini, C. Manferdini, K. Codeluppi, F. Grassi, A. Facchini and G. Lisignoli (2008). "Expression of CXC chemokines and their receptors is modulated during chondrogenic differentiation of human mesenchymal stem cells grown in three-dimensional scaffold: evidence in native cartilage." Tissue Engineering. Part A **14**(1): 97-105.

Darling, E. and K. Athanasiou (2005). "Rapid phenotypic changes in passaged articular chondrocyte subpopulations." J. Orthop. Res. **23**(2): 425-432.

De Ceuninck, F., C. Lesur, P. Pastoureau, A. Caliez and M. Sabatini (2004). Culture of Chondrocytes in Alginate Beads. Cartilage and Osteoarthritis, Vol. 1: Cellular and Molecular Tools. M. Sabatini, P. Pastoureau and F. De Ceuninck, Humana Press Inc. **100:** 15.

de Jonge, H., R. Fehrmann, E. de Bont, R. Hofstra, F. Gerbens, W. Kamps, E. de Vries, A. van der Zee, G. te Meerman and A. ter Elst (2007). "Evidence based selection of housekeeping genes." PloS One **2**(9): e898.

Debeb, B., L. Lacerda, W. Xu, R. Larson, T. Solley, R. Atkinson, E. Sulman, N. Ueno, S. Krishnamurthy, J. Reuben, T. Buchholz and W. Woodward (2012). "Histone deacetylase inhibitors stimulate dedifferentiation of human breast cancer

cells through WNT/β-catenin signaling." Stem Cells (Dayton, Ohio) **30**(11): 2366-2377.

Diaz-Romero, J., A. Quintin, E. Schoenholzer, C. Pauli, A. Despont, M. Zumstein, S. Kohl and D. Nesic (2014). "S100A1 and S100B expression patterns identify differentiation status of human articular chondrocytes." Journal of Cellular Physiology **229**(8): 1106-1117.

Dispersyn, G. D., L. Mesotten, B. Meuris, A. Maes, L. Mortelmans, W. Flameng, F. Ramaekers and M. Borgers (2002). "Dissociation of cardiomyocyte apoptosis and dedifferentiation in infarct border zones." European Heart Journal **23**(11): 849-857.

Dominici, M., K. Le Blanc, I. Mueller, I. Slaper-Cortenbach, F. Marini, D. Krause, R. Deans, A. Keating, D. Prockop and E. Horwitz (2006). "Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement." Cytotherapy **8**(4): 315-317.

Driesen, R., F. Verheyen, W. Debie, E. Blaauw, F. Babiker, R. Cornelussen, J. Ausma, M.-H. Lenders, M. Borgers, C. Chaponnier and F. Ramaekers (2009). "Re-expression of alpha skeletal actin as a marker for dedifferentiation in cardiac pathologies." Journal of Cellular and Molecular Medicine **13**(5): 896-908.

Dunning, M., M. Smith, M. Ritchie and S. Tavaré (2007). "beadarray: R classes and methods for Illumina bead-based data." Bioinformatics **23**(16): 2183-2184.

El-Gohary, Y., S. Tulachan, J. Wiersch, P. Guo, C. Welsh, K. Prasadan, J. Paredes, C. Shiota, X. Xiao, Y. Wada, M. Diaz and G. Gittes (2014). "A smad signaling network regulates islet cell proliferation." Diabetes **63**(1): 224-236.

Enver, T., M. Pera, C. Peterson and P. W. Andrews (2009). "Stem Cell States, Fates, and the Rules of Attraction." Cell Stem Cell **4**(5): 387-397.

Falcon, S. and R. Gentleman (2007). "Using GOstats to test gene lists for GO term association." Bioinformatics (Oxford, England) **23**(2): 257-258.

Ferrell, J. (2012). "Bistability, bifurcations, and Waddington's epigenetic landscape." Current biology : CB **22**(11): R458-R466.

Friedmann-Morvinski, D., E. Bushong, E. Ke, Y. Soda, T. Marumoto, O. Singer, M. Ellisman and I. Verma (2012). "Dedifferentiation of Neurons and Astrocytes by Oncogenes Can Induce Gliomas in Mice." Science **338**(6110): 1080-1084.

Gentleman, R., V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Yang and J. Zhang (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biology **5**(10): R80.

Gibson, M. A., M. L. Finnis, J. S. Kumaratilake and E. G. Cleary (1998). "Microfibril-associated glycoprotein-2 (MAGP-2) is specifically associated with fibrillin-containing microfibrils but exhibits more restricted patterns of tissue localization and developmental expression than its structural relative MAGP-1." The Journal of Histochemistry and Cytochemistry **46**(8): 871-886.

Gilsohn, E. and T. Volk (2010). "Fine tuning cellular recognition: The function of the leucine rich repeat (LRR) trans-membrane protein, LRT, in muscle targeting to tendon cells." Cell Adhesion & Migration **4**(3): 368-371.

Grafi, G. (2004). "How cells dedifferentiate: a lesson from plants." Developmental Biology **268**(1): 1-6.

Güngörmüş, C. and D. Kolankaya (2012). "Gene expression of tendon collagens and tenocyte markers in long-term monolayer and high-density cultures of rat tenocytes." Connective Tissue Research **53**(6): 485-491.

Halfon, S., N. Abramov, B. Grinblat and I. Ginis (2011). "Markers distinguishing mesenchymal stem cells from fibroblasts are downregulated with passaging." Stem Cells and Development **20**(1): 53-66.

Halper, J. and M. Kjaer (2014). "Basic components of connective tissues and extracellular matrix: elastin, fibrillin, fibulins, fibrinogen, fibronectin, laminin, tenascins and thrombospondins." Advances in Experimental Medicine and Biology **802**: 31-47.

Havis, E., M.-A. Bonnin, I. Olivera-Martinez, N. Nazaret, M. Ruggiu, J. Weibel, C. Durand, M.-J. Guerquin, C. Bonod-Bidaud, F. Ruggiero, R. Schweitzer and D.

Duprez (2014). "Transcriptomic analysis of mouse limb tendon cells during development." Development **141**(19): 3683-3696.

Hayflick, L. (1961). "The serial cultivation of human diploid cell strains." Experimental Cell Research **25**(3): 585-621.

Hinck, L. (2004). "The versatile roles of "axon guidance" cues in tissue morphogenesis." Developmental Cell **7**(6): 783-793.

Hoffmann, M., H. Bruns, L. Bäckdahl, P. Neregård, B. Niederreiter, M. Herrmann, A. I. Catrina, B. Agerberth and R. Holmdahl (2013). "The cathelicidins LL-37 and rCRAMP are associated with pathogenic events of arthritis in humans and rats." Annals of the Rheumatic Diseases **72**(7): 1239-1248.

Holmberg, J. and T. Perlmann (2012). "Maintaining differentiated cellular identity." Nat. Rev. Genet. **13**(6): 429-439.

Holmes, G. and L. Niswander (2001). "Expression of slit-2 and slit-3 during chick development." Dev. Dyn. **222**(2): 301-307.

Huang, D., B. Sherman and R. Lempicki (2008). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat. Protocols **4**(1): 44-57.

Huang, K. and L. D. Wu (2009). "YKL-40: a potential biomarker for osteoarthritis." The Journal of International Medical Research **37**(1): 18-24.

Huang, S. (2009). "Reprogramming cell fates: reconciling rarity with robustness." BioEssays **31**(5): 546-560.

Husson, F., Le, S. and Pages, J. (2010). Exploratory Multivariate Analysis by Example Using R, CRC Press.

Jopling, C., S. Boue and J. Belmonte (2011). "Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration." Nature Reviews Molecular Cell Biology **12**(2): 79-89.

Jopling, C., E. Sleep, M. Raya, M. Martí, A. Raya and J. C. Izpisúa Belmonte (2010). "Zebrafish heart regeneration occurs by cardiomyocyte dedifferentiation and proliferation." Nature **464**(7288): 606-609.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Research **28**(1): 27-30.

Kapacee, Z., S. Richardson, Y. Lu, T. Starborg, D. Holmes, K. Baar and K. Kadler (2008). "Tension is required for fibripositor formation." Matrix Biology **27**(4): 371-375.

Kapacee, Z., C.-Y. Yeung, Y. Lu, D. Crabtree, D. Holmes and K. Kadler (2010). "Synthesis of embryonic tendon-like tissue by human marrow stromal/mesenchymal stem cells requires a three-dimensional environment and transforming growth factor β 3." Matrix Biology **29**(8): 668-677.

Kolde, R., S. Laur, P. Adler and J. Vilo (2012). "Robust rank aggregation for gene list integration and meta-analysis." Bioinformatics **28**(4): 573-580.

Kragl, M., D. Knapp, E. Nacu, S. Khattak, M. Maden, H. H. Epperlein and E. Tanaka (2009). "Cells keep a memory of their tissue origin during axolotl limb regeneration." Nature **460**(7251): 60-65.

Krämer, A., J. Green, J. Pollard and S. Tugendreich (2014). "Causal analysis approaches in Ingenuity Pathway Analysis." Bioinformatics **30**(4): 523-530.

Kramer, S. G., T. Kidd, J. H. Simpson and C. S. Goodman (2001). "Switching repulsion to attraction: changing responses to slit during transition in mesoderm migration." Science (New York, N.Y.) **292**(5517): 737-740.

Lakemeier, S., J. Reichelt, T. Patzer, S. Fuchs-Winkelmann, J. Paletta and M. Schofer (2010). "The association between retraction of the torn rotator cuff and increasing expression of hypoxia inducible factor 1α and vascular endothelial growth factor expression: an immunohistological study." BMC Musculoskeletal Disorders **11**(1): 230.

Langfelder, P. and S. Horvath (2008). "WGCNA: an R package for weighted correlation network analysis." BMC Bioinformatics **9**(1): 559.

Lee, Y. J. and I. Baskakov (2010). "Treatment with normal prion protein delays differentiation and helps to maintain high proliferation activity in human embryonic stem cells." Journal of Neurochemistry **114**(2): 362-373.

Lehoczky, J., B. Robert and C. Tabin (2011). "Mouse digit tip regeneration is mediated by fate-restricted progenitor cells." Proceedings of the National Academy of Sciences of the United States of America **108**(51): 20609-20614.

Lejard, V., F. Blais, M.-J. Guerquin, A. Bonnet, M.-A. Bonnin, E. Havis, M. Malbouyres, C. B. Bidaud, G. Maro, P. Gilardi-Hebenstreit, J. Rossert, F. Ruggiero and D. Duprez (2011). "EGR1 and EGR2 involvement in vertebrate tendon differentiation." The Journal of Biological Chemistry **286**(7): 5855-5867.

Leys, C., C. Ley, O. Klein, P. Bernard and L. Licata (2013). "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median." Journal of Experimental Social Psychology **49**(4): 764-766.

Linhardt, R. J. (2001). "Analysis of glycosaminoglycans with polysaccharide lyases." Current protocols in molecular biology / edited by Frederick M. Ausubel **Chapter 17**: 17.13B.11-17.13B.16.

Liu, H., S. Zhu, C. Zhang, P. Lu, J. Hu, Z. Yin, Y. Ma, X. Chen and H. OuYang (2014). "Crucial transcription factors in tendon development and differentiation: their potential for tendon regeneration." Cell and Tissue Research **356**(2): 287-298.

Liu, J., L. Zhang, D. Wang, H. Shen, M. Jiang, P. Mei, P. Hayden, J. Sedor and H. Hu (2003). "Congenital diaphragmatic hernia, kidney agenesis and cardiac defects associated with Slit3-deficiency in mice." Mechanisms of Development **120**(9): 1059-1070.

Livak, K. and T. Schmittgen (2001). "Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta}$CT Method." Methods **25**(4): 402-408.

Ma, B., J. C. Leijten, L. Wu, M. Kip, C. A. van Blitterswijk, J. N. Post and M. Karperien (2013). "Gene expression profiling of dedifferentiated human articular chondrocytes in monolayer culture." Osteoarthritis and Cartilage **21**(4): 599-603.

Maleki, M., F. Ghanbarvand, M. Reza Behvarz, M. Ejtemaei and E. Ghadirkhomi (2014). "Comparison of Mesenchymal Stem Cell Markers in Multiple Human Adult Stem Cells." International Journal of Stem Cells **7**(2): 118-126.

Mapp, P. I., P. S. Avery, D. F. McWilliams, J. Bowyer, C. Day, S. Moores, R. Webster and D. A. Walsh (2008). "Angiogenesis in two animal models of osteoarthritis." Osteoarthritis and Cartilage **16**(1): 61-69.

Marchuk, L., P. Sciore, C. Reno, C. B. Frank and D. A. Hart (1998). "Postmortem stability of total RNA isolated from rabbit ligament, tendon and cartilage." Biochimica et Biophysica Acta **1379**(2): 171-177.

Marcil, A., E. Dumontier, M. Chamberland, S. Camper and J. Drouin (2003). "Pitx1 and Pitx2 are required for development of hindlimb buds." Development (Cambridge, England) **130**(1): 45-55.

Massam-Wu, T., M. Chiu, R. Choudhury, S. Chaudhry, A. Baldwin, A. McGovern, C. Baldock, A. Shuttleworth and C. Kielty (2010). "Assembly of fibrillin microfibrils governs extracellular deposition of latent TGFβ." Journal of Cell Science **123**(17): 3006-3018.

Meloty-Kapella, C., M. Degen, R. Chiquet-Ehrismann and R. Tucker (2006). "Avian tenascin-W: expression in smooth muscle and bone, and effects on calvarial cell spreading and adhesion in vitro." Developmental dynamics **235**(6): 1532-1542.

Millar, N., J. Reilly, S. Kerr, A. Campbell, K. Little, W. Leach, B. Rooney, G. Murrell and I. McInnes (2012). "Hypoxia: a critical regulator of early human tendinopathy." Annals of the Rheumatic Diseases **71**(2): 302-310.

Miranda, A., E. Pericuesta, M. Á. Ramírez and A. Gutierrez-Adan (2011). "Prion protein expression regulates embryonic stem cell pluripotency and differentiation." PloS One **6**(4): e18422.

Murtagh, F. (1985). Multidimensional Clustering Algorithms, Wuerzburg: Physica-Verlag.

Nagoshi, N., S. Shibata, M. Hamanoue, Y. Mabuchi, Y. Matsuzaki, Y. Toyama, M. Nakamura and H. Okano (2011). "Schwann cell plasticity after spinal cord injury shown by neural crest lineage tracing." Glia **59**(5): 771-784.

Ochi, K., A. Derfoul and R. Tuan (2006). "A predominantly articular cartilage-associated gene, SCRG1, is induced by glucocorticoid and stimulates chondrogenesis in vitro." Osteoarthritis and Cartilage **14**(1): 30-38.

Oliphant, A., D. Barker, J. Stuelpnagel and M. Chee (2002). "BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping." BioTechniques **32**: S56-S61.

Perkins, J., J. Dawes, S. McMahon, D. Bennett, C. Orengo and M. Kohl (2012). "ReadqPCR and NormqPCR: R packages for the reading, quality checking and normalisation of RT-qPCR quantification cycle (Cq) data." BMC Genomics **13**(1): 296.

Porrello, E., A. Mahmoud, E. Simpson, J. Hill, J. Richardson, E. Olson and H. Sadek (2011). "Transient regenerative potential of the neonatal mouse heart." Science (New York, N.Y.) **331**(6020): 1078-1080.

Poss, K. (2010). "Advances in understanding tissue regenerative capacity and mechanisms in animals." Nature Reviews Genetics **11**(10): 710-722.

Prasad, A., S. Kumar, C. Dessimoz, S. Bleuler, O. Laule, T. Hruz, W. Gruissem and P. Zimmermann (2013). "Global regulatory architecture of human, mouse and rat tissue transcriptomes." BMC Genomics **14**(1): 716.

Ramirez, F. and D. Rifkin (2009). "Extracellular microfibrils: contextual platforms for TGFbeta and BMP signaling." Current Opinion in Cell Biology **21**(5): 616-622.

Rando, Thomas A. and Howard Y. Chang (2012). "Aging, Rejuvenation, and Epigenetic Reprogramming: Resetting the Aging Clock." Cell **148**(1): 46-57.

Rinn, J. L., C. Bondre, H. B. Gladstone, P. O. Brown and H. Y. Chang (2006). "Anatomic Demarcation by Positional Variation in Fibroblast Gene Expression Programs." PLoS Genetics **2**(7): e119.

Ritty, T., K. Ditsios and B. Starcher (2002). "Distribution of the elastic fiber and associated proteins in flexor tendon reflects function." Anat. Rec. **268**(4): 430-440.

Roach, H., G. Mehta, R. Oreffo, N. Clarke and C. Cooper (2003). "Temporal analysis of rat growth plates: cessation of growth with age despite presence of a physis." The Journal of Histochemistry and Cytochemistry **51**(3): 373-383.

Russ, H., E. Sintov, L. Anker-Kitai, O. Friedman, A. Lenz, G. Toren, C. Farhy, M. Pasmanik-Chor, V. Oron-Karni, P. Ravassard and S. Efrat (2011). "Insulin-producing cells generated from dedifferentiated human pancreatic beta cells expanded in vitro." PloS One **6**(9): e25566.

Sagi, B., P. Maraghechi, V. S. Urban, B. Hegyi, A. Szigeti, R. Fajka-Boja, G. Kudlik, K. Nemet, E. Monostori, E. Gocza and F. Uher (2012). "Positional identity of murine mesenchymal stem cells resident in different organs is determined in the postsegmentation mesoderm." Stem Cells Dev **21**(5): 814-828.

Saito, T., T. Ikeda, K. Nakamura, U.-i. Chung and H. Kawaguchi (2007). "S100A1 and S100B, transcriptional targets of SOX trio, inhibit terminal differentiation of chondrocytes." EMBO Reports **8**(5): 504-509.

Sandberg, R. and I. Ernberg (2005). "The molecular portrait of in vitro growth by meta-analysis of gene-expression profiles." Genome Biology **6**(8).

Sandoval-Guzmán, T., H. Wang, S. Khattak, M. Schuez, K. Roensch, E. Nacu, A. Tazaki, A. Joven, E. Tanaka and A. Simon (2014). "Fundamental Differences in Dedifferentiation and Stem Cell Recruitment during Skeletal Muscle Regeneration in Two Salamander Species." Cell Stem Cell **14**(2): 174-187.

Schmittgen, T. and K. Livak (2008). "Analyzing real-time PCR data by the comparative C(T) method." Nature Protocols **3**(6): 1101-1108.

Schulze-Tanzil, G. (2009). "Activation and dedifferentiation of chondrocytes: Implications in cartilage injury and repair." Annals of Anatomy - Anatomischer Anzeiger **191**(4): 325-338.

Schweitzer, R., J. Chyung, L. Murtaugh, A. Brent, V. Rosen, E. Olson, A. Lassar and C. Tabin (2001). "Analysis of the tendon cell fate using Scleraxis, a specific marker for tendons and ligaments." Development **128**(19): 3855-3866.

Schweitzer, R., E. Zelzer and T. Volk (2010). "Connecting muscles to tendons: tendons and musculoskeletal development in flies and vertebrates." Development (Cambridge, England) **137**(17): 2807-2817.

Schwitalla, S., Alexander A. Fingerle, P. Cammareri, T. Nebelsiek, Serkan I. Göktuna, Paul K. Ziegler, O. Canli, J. Heijmans, David J. Huels, G. Moreaux, Rudolf A. Rupec, M. Gerhard, R. Schmid, N. Barker, H. Clevers, R. Lang, J. Neumann, T. Kirchner, Makoto M. Taketo, Gijs R. van den Brink, Owen J. Sansom, Melek C. Arkan and Florian R. Greten (2013). "Intestinal Tumorigenesis Initiated by Dedifferentiation and Acquisition of Stem-Cell-like Properties." Cell **152**(1–2): 25-38.

Smyth, G. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Statistical Applications in Genetics and Molecular Biology **3**(1): Article 3.

Sugimoto, K., S. Gordon and E. Meyerowitz (2011). "Regeneration in plants and animals: dedifferentiation, transdifferentiation, or just differentiation?" Trends in Cell Biology **21**(4): 212-218.

Supek, F., M. Bošnjak, N. Škunca and T. Šmuc (2011). "REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms." PLoS One **6**(7): e21800.

Szibor, M., J. Pöling, H. Warnecke, T. Kubin and T. Braun (2014). "Remodeling and dedifferentiation of adult cardiomyocytes during disease and regeneration." Cellular and Molecular Life Sciences **71**(10): 1907-1916.

Takahashi, K. and S. Yamanaka (2006). "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." Cell **126**(4): 663-676.

Takeuchi, J., K. Koshiba-Takeuchi, K. Matsumoto, A. Vogel-Hopker, M. Naitoh-Matsuo, K. Ogura, N. Takahashi, K. Yasuda and T. Ogura (1999). "Tbx5 and

Tbx4 genes determine the wing/leg identity of limb buds." Nature **398**(6730): 810-814.

Talchai, C., S. Xuan, Hua V. Lin, L. Sussel and D. Accili (2012). "Pancreatic β Cell Dedifferentiation as a Mechanism of Diabetic β Cell Failure." Cell **150**(6): 1223-1234.

Tarca, A. L., S. Draghici, P. Khatri, S. Hassan, P. Mittal, J.-S. Kim, C. J. Kim, J. P. Kusanovic and R. Romero (2009). "A novel signaling pathway impact analysis." Bioinformatics (Oxford, England) **25**(1): 75-82.

Tata, P. R., H. Mou, A. Pardo-Saganta, R. Zhao, M. Prabhu, B. Law, V. Vinarsky, J. Cho, S. Breton, A. Sahay, B. Medoff and J. Rajagopal (2013). "Dedifferentiation of committed epithelial cells into stem cells in vivo." Nature **503**(7475): 218-223.

Tchetina, E. (2011). "Developmental mechanisms in articular cartilage degradation in osteoarthritis." Arthritis **2011**: Article ID: 683970.

Trivedi, C. M., T. P. Cappola, K. B. Margulies and J. A. Epstein (2011). "Homeodomain Only Protein X is down-regulated in human heart failure." Journal of Molecular and Cellular Cardiology **50**(6): 1056-1058.

Tucker, R. P. and R. Chiquet-Ehrismann (2009). "The regulation of tenascin expression by tissue microenvironments." Biochimica et Biophysica Acta (BBA) - Molecular Cell Research **1793**(5): 888-892.

Van Gelder, R. N., M. E. von Zastrow, A. Yool, W. C. Dement, J. D. Barchas and J. H. Eberwine (1990). "Amplified RNA synthesized from limited quantities of heterogeneous cDNA." Proceedings of the National Academy of Sciences of the United States of America **87**(5): 1663-1667.

Vandesompele, J., K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe and F. Speleman (2002). "Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes." Genome Biology **3**(7).

Vargesson, N., V. Luria, I. Messina, L. Erskine and E. Laufer (2001). "Expression patterns of Slit and Robo family members during vertebrate limb development." Mechanisms of Development **106**(1-2): 175-180.

Volk, T. (1999). "Singling out Drosophila tendon cells: a dialogue between two distinct cell types." Trends in Genetics **15**(11): 448-453.

Volohonsky, G., G. Edenfeld, C. Klämbt and T. Volk (2007). "Muscle-dependent maturation of tendon cells is induced by post-transcriptional regulation of stripeA." Development **134**(2): 347-356.

Von der Mark, K., V. Gauss, H. Von der Mark and P. Muller (1977). "Relationship between cell shape and type of collagen synthesised as chondrocytes lose their cartilage phenotype in culture." Nature **267**(5611): 531-532.

Wang, K. C., J. A. Helms and H. Y. Chang (2009). "Regeneration, repair and remembering identity: the three Rs of Hox gene expression." Trends in Cell Biology **19**(6): 268-275.

Wenke, A. K., S. Niebler, S. Grässel and A. K. Bosserhoff (2011). "The transcription factor AP-2ε regulates CXCL1 during cartilage development and in osteoarthritis." Osteoarthritis and Cartilage **19**(2): 206-212.

Yammani, R. (2012). "S100 proteins in cartilage: role in arthritis." Biochimica et Biophysica Acta **1822**(4): 600-606.

Ye, J., G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen and T. Madden (2012). "Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction." BMC Bioinformatics **13**(1): 134.

Ye, Q., G. Zünd, P. Benedikt, S. Jockenhoevel, S. Hoerstrup, S. Sakyama, J. Hubbell and M. Turina (2000). "Fibrin gel as a three dimensional matrix in cardiovascular tissue engineering." European Journal of Cardio-Thoracic Surgery **17**(5): 587-591.

Zaitseva, M., B. Vollenhoven and P. Rogers (2006). "In vitro culture significantly alters gene expression profiles and reduces differences between myometrial and fibroid smooth muscle cells." Molecular Human Reproduction **12**(3): 187-207.

Zakany, J. and D. Duboule (2007). "The role of Hox genes during vertebrate limb development." Pattern Formation and Developmental Mechanisms **17**(4): 359-366.

Zhang, Y., T.-S. Li, S.-T. Lee, K. Wawrowsky, K. Cheng, G. Galang, K. Malliaras, R. Abraham, C. Wang and E. Marbán (2010). "Dedifferentiation and proliferation of mammalian cardiomyocytes." PloS One **5**(9): e12559.

Zhou, J. and S. Huang (2011). "Understanding gene circuits at cell-fate branch points for rational cell reprogramming." Trends in Genetics **27**(2): 55-62.

R Packages

No associated peer-reviewed publications

FactoMineR: Husson, F., Le, S. and Pages, J. (2010). Exploratory Multivariate Analysis by Example Using R, *Chapman and Hall*. Version 1.25 (2013).

org.Rn.eg.db. (Pages H, Carlson M, Falcon S and Li N. AnnotationDbi: Annotation Database Interface. R package version 1.28.1(2014).

Appendix 1

R Codes for Illumina microarray analysis

Pre-processing, normalisation and differential expression analysis

```
#series of 40 microarrays are read individually.  Files contain
raw Bead Level Data (BLData).
setwd()
library(beadarray)
library(illuminaRatv1.db) #correct annotation package


    BLData.arrayName=readIllumina(
    useImages=FALSE,illuminaAnnotation="Ratv1") #change name
    for each file
```

```
#save all loaded raw data text files
#check what has been read
slotNames(BLData. BLData.arrayName)
#first ten rows of data from all columns
BLData.arrayName [[1]][1:10,]
#boxplot data
boxplot(BLData.arrayName,las=2,outline=FALSE,ylim=c(4,12))
#apply BASH algorithm - this takes a while for each array!
BLData.arrayName.bsh
=BASH(BLData.arrayName,array=1,useLocs=FALSE)


#save a separate file of only .bsh files so that these may
be accessed.  Remove all raw files prior to saving .bsh
files


rm(BLData.arrayName,…)


#set weights derived from BASH assessment of beads.  This
needs to be done for each array individually
BLData.arrayName=setWeights(
BLData.arrayName,wts=BLData.arrayName.bsh$wts,array=1,
combine=FALSE,wtName='wts')
#add quality information from BASH to bead level data for
each array individually
BLData.arrayName=insertSectionData(BLData.arrayName,what="B
ASHQC",data=BLData.arrayName.bsh$QC)
#check that extended score etc have been added
BLData.arrayName@sectionData
#plots positive control for housekeeping or biotin
poscontPlot(BLData.arrayName)
png('controlplots.png')
#a general control plot for all data
combinedControlPlot(BLData.arrayName)
dev.off()
#combine all arrays in a single expression set - this has
to be done one-by-one in beadarray
BLData=combine(BLData.arrayName1,BLData.arrayName2)
BLData=combine(BLData,BLData.arrayName3)
#continue for the other arrays adding to BLData
#Summarise probe data
myMean=function(x)mean(x,na.rm=TRUE)
mySd=function(x)sd(x,na.rm=TRUE)
greenChannel=new("illuminaChannel",greenChannelTransform,il
luminaOutlierMethod,myMedian,myMad,"Grn")
BSData=summarize(BLData,list(greenChannel),
useSampleFac=TRUE,sampleFac=rep(1:36,each=1),
weightNames="wts",removeUnMappedProbes=TRUE);
det=calculateDetection(BSData,status=fData(BSData)$Status,
negativeLabel="negative")
Detection(BSData)=det
#Transform and normalise across arrays.  Both quantile and
loess strategies are shown
#QUANTILE
BSData.q=normaliseIllumina(BSData,method="quantile",transfo
rm="log2")
```

```
     #LOESS

library(limma)

     BSData=normaliseIllumina(
     BSData,method="none",transform="log2")
     BSData.loess<-normalizeCyclicLoess(exprs(BSData),
     weights = NULL, span=0.7, iterations = 3, method = "affy")
     ##no longer an eSet, just a normalised matrix
     write.csv(BSData.loess,file="BSData_loess.csv")


     #filter probes
     BSData.genes=BSData.q[
     which(fData(BSData)$Status=="regular"), ]
     expressed=apply(Detection(BSData.genes)<0.05,1,any)
     BSData.filt=BSData.genes[expressed,]
     ###export BSData.filt to WGCNA and GOstats
     ############filter
     ID=as.character(featureNames(BSData.q))
     #addFeatureData
     qual=unlist(mget(
     ID,illuminaRatv1PROBEQUALITY,ifnotfound=NA))
     table(qual)
       rem<- qual == "No match" | qual == "Bad" | is.na(qual)
     #vector of probes to be removed
     BSData.filt=BSData.q[!rem, ]
     dim(BSData.filt)

     #Matrix design for differential expression analysis for 36
     arrays.  Assign abbreviated names to each array based on
     sample group and replicate
     design<model.matrix(~0+factor(
     c(1,1,1,1,2,2,3,3,3,3,3,3,3,4,4,4,5,5,6,6,6,6,6,6,6,6,7,7
     ,7,7,8,8,8,8,8)))
     colnames(design)<-
     c("dC_ALG","dC_FIB","dC","dF","dT_FIB","dT","nC","nT")
     rownames(design)<c("dC_ALG3","dC_ALG4","dC_ALG1","dC_ALG2",
     "dC_FIB1","dC_FIB2","dC1","dC2","dC3","dC4","dC5","dC6","dC
     7","dC8","dF1","dF2","dF3","dT_FIB1","dT_FIB2","dT1","dT2",
     "dT3","dT4","dT5","dT6","dT7","dT8","nC2","nC3","nC4","nC5"
     ,"nT1","nT2","nT3","nT4","nT5")

     ###Differential expression and feature data for loess
     normalised matrix
     ID<-rownames(BSData.loess)
     symbol=mget(ID,illuminaRatv1SYMBOL,ifnotfound=NA)
     genename=mget(ID,illuminaRatv1GENENAME,ifnotfound=NA)
     entrezID=mget(ID,illuminaRatv1ENTREZID,ifnotfound=NA)
     anno=data.frame(Illumina_ID=ID,Symbol=as.character(symbol),
     EntrezID=as.numeric(entrezID),
     GeneName=as.character(genename))
     fit<-lmFit(ajm.loess,design)
     contrast.matrix<-makeContrasts(dC-dT,levels=design)
     #set up the matrix and then you can include or exclude the
     samples that you want
     fit2<-contrasts.fit(fit,contrast.matrix)
     fit2<-eBayes(fit2)
     fit2$gene=anno
     rankCresults=topTable(
     fit2,coef=1,number=1500,lfc=1.4,adjust.method="fdr",
```

```
sort.by="logFC",genelist=fit2$gene)
#export results to working directory in ranked format
write.table(rankCresults,file="rankCresults.txt",sep="\t")

###########
aw=arrayWeights(exprs(BSData.filt),design)
#Differential expression analysis for quantile normalised
data
fit<-lmFit(exprs(BSData.filt),design, weights=aw)
ID=featureNames(BSData.filt)
chr=mget(ID,illuminaRatv1CHR,ifnotfound=NA)
refseq=mget(ID,illuminaRatv1REFSEQ,ifnotfound=NA)
entrezID=mget(ID,illuminaRatv1ENTREZID,ifnotfound=NA)
symbol=mget(ID,illuminaRatv1SYMBOL,ifnotfound=NA)
genename=mget(ID,illuminaRatv1GENENAME,ifnotfound=NA)
probequality=mget(ID,illuminaRatv1PROBEQUALITY,
ifnotfound=NA)
GO=mget(ID,illuminaRatv1GO,ifnotfound=NA)

anno=data.frame(
Illumina_ID=ID,Chr=as.character(chr),
RefSeq=as.character(refseq),
EntrezID=as.numeric(entrezID),
Symbol=as.character(symbol),
GeneName=as.character(genename),
ProbeQuality=as.character(probequality),
GOterm=as.character(GO))
#linear model fit and contrast matrix
fit<-lmFit(exprs(BSData.filt),design)
contrast.matrix<-makeContrasts(dC_ALG-dC,levels=design)
#set up the matrix and then you can include or exclude the
samples that you want
fit2<-contrasts.fit(fit,contrast.matrix)
fit2<-eBayes(fit2)
fit2$gene=anno
rankCresults=topTable(
fit2,coef=1,number=1500,lfc=1.4,adjust.method="fdr",
sort.by="logFC",genelist=fit2$gene)
write.table(rankCresults,file="rankCresults.txt",sep="\t")

#Hierarchical clustering of quantile normalised data
d=dist(t(exprs(BSData.q)))
plot(hclust(d))
```

Hierarchical clustering and heatmap

```
setwd("/Users/XXX")
data<-read.csv("BSData_loess.csv",header=TRUE)
colnames(data)[1]<-'IlluminaID'

ArrayName=names(data.frame(data[,-1]))
GeneName=data$EntrezID
exprs=data.frame(t(data[,-1]))
names(exprs)=data[,1]
dimnames(exprs)[[1]]=names(data.frame(data[,-1]))
exprs.v=as.vector(apply(as.matrix(exprs),2,var,na.rm=T))
keep=exprs.v>0.8

library(WGCNA)

filt=exprs[,keep]
```

```
   GeneName=GeneName[keep]
   ADJ1=abs(cor(filt,use="p"))^9 #create adjacency matrix
   k=as.vector(apply(ADJ1,2,sum, na.rm=T))
   datExpr=filt[, rank(-k,ties.method="first" )<=500]
   rename<-t(datExpr)
   colnames(rename)<-
   c(rep("3D",6),rep("2D",11),rep("3D",2),rep("2D",8),
   rep("Cartilage",4),rep("Tendon",5))
   map<-as.matrix(rename)

   #Define and export the heatmap groups

library(gplots)
library(RColorBrewer)

   hm <- heatmap.2(map)
   hc <- as.hclust(hm$rowDendrogram)
   #define the height at which the dendrogram is cut
   groups<-cutree(hc, h=25) [hc$order]
   names<-names(groups)
   groups1<-unname(groups)
   groups2<-data.frame("Symbol"=names,"Groups"=groups1)
   write.csv(groups2,file="heatmapGroups.csv",row.names=FALSE)

   ##Create heatmap with the row groups and columns colour-coded
   groups<-cutree(hc,h=25)
   cols <- brewer.pal(max(groups), "Set3")
   setwd("")
   pdf(file = "Illumina_heatmap2.pdf", width= 8,
   height = 8,useDingbats=F)
   par(oma=c(2,2,2,2))

   heatmap.2(map,scale="row",col=greenred(100),
   colsep=c(4,9,17),sepcolor="white",sepwidth=c(0.1,0.1),
   trace="none",density.info="none",RowSideColors=cols[groups],
   ColSideColors=c(rep("firebrick1",6),rep("midnightblue",11),
   rep("firebrick1",2),rep("midnightblue",8),
   rep("lightsteelblue3",4),rep("goldenrod2",5)),
   cexRow=0.07,cexCol=1)

   dev.off()
   ##retain 'map', a matrix of gene expression values, for PCA
```

Principal Component Analysis

```
   #PCA for 36 arrays filtered on covariance

library(FactoMineR)
library(RColorBrewer)

   #re-order columns from heatmap matrix so that they lie: 2D,
   #3D, native
   map2<-
   map[,c(7,8,9,10,11,12,13,14,15,16,17,20,21,22,23,24,25,26,27,
   1,2,3,4,5,6,18,19,28,29,30,31,32,33,34,35,36)];
   res.pca<-PCA(t(map2),graph=FALSE,axes=c(1,2))

   PC1 <- res.pca$ind$coord[,1]
   PC2 <- res.pca$ind$coord[,2]
```

```
   #define factors
   cell.type<-c("chondrocytes", "chondrocytes", "chondrocytes",
   "chondrocytes", "chondrocytes", "chondrocytes",
   "chondrocytes", "chondrocytes", "fibroblasts","fibroblasts",
   "fibroblasts","tenocytes","tenocytes", "tenocytes",
   "tenocytes", "tenocytes", "tenocytes","tenocytes",
   "tenocytes","chondrocytes", "chondrocytes", "chondrocytes",
   "chondrocytes","chondrocytes","chondrocytes","tenocytes",
   "tenocytes","chondrocytes", "chondrocytes", "chondrocytes",
   "chondrocytes", "tenocytes", "tenocytes", "tenocytes",
   "tenocytes", "tenocytes")
   cell.type<-as.data.frame(cell.type)
   condition<-
   c(rep("monolayer",19),rep("model.3D",8),rep("cartilage",4),re
   p("tendon",5))
   condition<-as.data.frame(condition)


   PCs <- data.frame(cbind(PC1,PC2,cell.type,condition))
   PCA.comp1<-res.pca$eig[1,2]
   PCA.comp2<-res.pca$eig[2,2]

   #Colours for plot
   mypalette<-c("gray0","gray88","gray64","gray40")
   #Prepare and export PCA plot

library(ggplot2)

   setwd("/Users/XXX")
   pdf(file = "Illumina_PCA_Figure.pdf", width=8,
   height=8,useDingbats=F)
   par(mar=c(1,1,1,1))
   p<-ggplot(PCs)
   p<-p+geom_point(aes(PC1,PC2,color=condition,shape=cell.type),
   size=6,alpha=0.6)+
   scale_colour_manual(values=mypalette)+
   labs(list(x=sprintf("PC1(%.1f%%)",PCA.comp1),
   y=sprintf("PC2(%.1f%%)",PCA.comp2)))+
   theme_minimal(base_size=10,base_family="Helvetica")+
   theme(legend.position = c(.85,.7),text =
   element_text(size=12),plot.title=element_text(
   lineheight=.8,face="bold"))+
   ggtitle("Principal Component Analysis")+
   scale_shape_discrete(solid=T)
   p

   dev.off()
```

Hypergeometric testing of gene ontology analysis

```
#####Define the universe - all the genes on the Illumina
microarray.  Remove duplicate entries from Entrez annotations
##Whole chip
   setwd("/Users/… ")
   universe<-
   read.csv("RatRefv1.csv",header=TRUE,sep=",",as.is=TRUE)
   rem.dups<-universe[!duplicated(universe$Entrez_Gene_ID),]
   universe.entrez<-as.vector(rem.dups$Entrez_Gene_ID)
   length(universe.entrez)
   table(is.na(universe.entrez))
```

```
  rem.universe<-universe.entrez=="NA"|is.na(universe.entrez)
  filt<-universe.entrez[!rem.universe]
  UNIVERSE<-as.numeric(filt)

  #setwd("/Users/……. ")
  #read in differential expression lists of genes by Entrez ID
  to functionally annotate
  nt.up<-read.csv("working.csv",header=TRUE)
  nt.entrez<-nt.up$Entrez
  rem.NA<-nt.entrez=="NA"|is.na(nt.entrez)
  table(rem.NA)
  #filt<-nt.entrez[!rem.NA]
  dups<-duplicated(nt.entrez)
  table(dups)

  no.dups=nt.entrez[!dups]
  nt.final<-as.numeric(no.dups)

library(illuminaRatv1.db)
library(GOstats)

  hgCutoff <- 0.001  #statistical cut-off

  #Perform each in turn for biological process, cellular
  compartment and metabolic function.  Change name of output
  file on each occasion.
  params <-
  new("GOHyperGParams",geneIds=nt.final,universeGeneIds=UNIVERS
  E,annotation="illuminaRatv1.db",ontology="BP",pvalueCutoff=hg
  Cutoff,conditional=FALSE,testDirection="over")

  #params <-
  new("GOHyperGParams",geneIds=nt.final,universeGeneIds=UNIVERS
  E,annotation="illuminaRatv1.db",ontology="MF",pvalueCutoff=hg
  Cutoff,conditional=FALSE,testDirection="over")

  #params <-
  new("GOHyperGParams",geneIds=nt.final,universeGeneIds=UNIVERS
  E,annotation="illuminaRatv1.db",ontology="CC",pvalueCutoff=hg
  Cutoff,conditional=FALSE,testDirection="over")

  hgOver <- hyperGTest(params)
  df=summary(hgOver,htmlLinks=FALSE) #TRUE returns links to
  AmiGO
  hgOver

  p.value<-df$Pvalue
  adjusted.p<-p.adjust(p.value,method="fdr")
  df$adj.Pvalue<-adjusted.p

   write.csv(file='GO.csv',df,row.names=FALSE)
```

SPIA pathway topology analysis

```
#Requires XML files to be downloaded from KEGG and stored within
a named folder within the same directory;
```

```
library(SPIA)

  setwd("/Volumes/XXX/SPIA")
  makeSPIAdata(kgml.path="/Volumes/XXX/kegg",organism="rno",out
  .path="/Volumes/XXX/kegg")
  #read in lists of differentially expressed genes as Entrez
  IDs
  top<-read.csv("SPIA.csv",header=TRUE,sep=",")
  setwd("/Users/alanmueller/Desktop/Thesis")
  #create universe based on microarray probes
  universe<-
  read.csv("RatRefv1.csv",header=TRUE,sep=",",as.is=TRUE)

  #ensure that everything in universe is found in top
  merged<-merge(top,universe,by.x<-
  "EntrezID",by.y="Entrez_Gene_ID")

  dim(merged)
  #[1] 2007 37
  top<-merged[!duplicated(merged$EntrezID),]
  dim(top)
  #[1] 1842 37
  top<-top[top$adj.P.Val<0.01,]
  dim(top)
  #[1] 1658 37

  de<-as.vector(top$log2FC)
  names(de)<-as.vector(top$EntrezID)
  head(de)

  dim(universe)
  #[1] 23405 28
  rem.dups<-universe[!duplicated(universe$Entrez_Gene_ID),]
  dim(rem.dups)
  #[1] 21494 28

  universe.entrez<-as.vector(rem.dups$Entrez_Gene_ID)

  #The SPIA algorithm takes as input the two vectors above and
  produces a table of pathways ranked from the most to the
  least significant.
  res<-spia(de=de, all=universe.entrez,
  organism="rno",data.dir="/Volumes/XXX/kegg/",nB=2000,
  plots=FALSE)
  #show first 15 pathways, omit KEGG links
  res[1:20,-12]
  plotP(res,threshold=0.05)

  setwd("/Users/XXX/SPIA_pathways")
  pdf("SPIA_pathways_SPIA.pdf")
  plotP(res,threshold=0.05)
  dev.off()
  results<-res[1:20,]
  write.table(results,file="SPIA_pathways_SPIA.txt")
```

3: A pathway topology approach predicts involvement of the PI-3K/Akt signalling pathway in musculoskeletal cell differentiation

Abstract

The systems biology approach that has been defined in this thesis demands iterative perturbation and quantification of network responses to rationally model a system. For comparisons to be relevant inherent differences in matricellular components between musculoskeletal tissues require to be resolved if common mechanistic processes are to be revealed. Defining consistent alterations in common prevailing regulatory mechanisms induced by the transition from *in vivo* to *in vitro* culture conditions for cartilage and tendon may be obscured by differences in sample complexity and composition. To extract evidence-based targets for future intervention studies depletion of extra- and peri-cellular matrix of cartilage and tendon cells was employed to reduce the complexity of samples and act as a system perturbation.

Matrix-depleted cells exhibited strong clustering by condition rather than cell type including co-clustering of native cartilage and tendon cells. High overlap in expression profiles and functional annotations for different cell types in the same condition suggested a common regulatory mechanism underlying de- and re-differentiation for both chondrocytes and tenocytes. Supporting previous results musculoskeletal development-associated genes including *Scx* (scleraxis), *Mkx*

(mohawk) and *Mustn1* (musculoskeletal, embryonic nuclear protein 1) were all more highly expressed in monolayer cells from both tissues. Defining consensus genes across microarray data sets isolated tissue-associated genes from those arising from other components of heterogenous samples providing putative tendon-associated genes including *Serpinf1* and *Mfap5*.

The PI-3K/Akt signalling pathway is a core signal transduction mechanism with multiple integrated roles including cellular differentiation. Results implicate the PI-3K/Akt signalling pathway as a key regulator of chondrocyte and tenocyte de- and re-differentiation using a pathway topology approach. Expression profiles were supportive of active PI-3K signalling in native cells and inhibition in monolayer cultures for both cell types indicative of a common regulatory pathway in de- and re-differentiation.

3.1: Introduction

3.1.1: Study rationale

Several concerns relating to the study design presented in **Chapter 2** encouraged the development of a second transcriptomic survey of chondrocytes and tenocytes from different environmental contexts.

Microarray platform

The Illumina RatRef-12 v1.0 Gene Expression BeadChip array, unlike, the human or mice equivalents from the same manufacturer, had not been re-annotated since the first market release; the omission of key musculoskeletal developmental regulators from the gene manifest for this platform (*Sox9* and *Scx*) was evident.

Furthermore, this platform was discontinued in December 2011 negating the possibility of additional samples being surveyed to equalize groups sizes, for example for fibrin constructs.

Reduction of study and sample complexity

Concerns arose with regard to the comparison of native cartilage and tendon to isolated cells in culture used in **Chapter 2** due to the inherent difference in the complexity of native tissue relative to monolayer culture and the relevance of comparing tissues of distinct composition. Gene expression profiles from native samples were also likely to be confounded by heterogenous cell populations, contaminants, and complex matrix components reduced in culture conditions.

Evidence in **Chapter 2** demonstrated that dedifferentiated cells developed a convergent gene expression profile when passaged in culture. There was support for the statement that cells in monolayer and in three-dimensional culture systems had expression profiles that were more akin to developmental and reparative profiles than to their tissues of origin. However, there was no consistent prediction for perturbed signalling pathways limiting rationale identification of the prevailing regulatory signalling mechanisms for de- and re-differentiation. This either suggested that the mechanisms of de- and re-differentiation were in fact different between cartilage and tendon or that the common regulatory mechanisms were obfuscated by the inherent differences in the complexity of each condition, differences in sample handling, and the presence of heterogenous cell populations.

In proteomic studies reduction of sample complexity, for example by fractionation (Cox and Emili 2006) or depletion (Fonslow, Stein et al. 2013), is encouraged.

Abundant matrix proteins can swamp mass spectrometry analysis and greatly limit the depth, or coverage, of a proteomic survey (Wilson, Whitelock et al. 2009). As part of a parallel proteomics study, presented in **Chapter 6**, a strategy to reduce sample complexity was devised. To ensure comparability between studies this strategy was extended to samples for transcriptomic survey. Globin (Mastrokolias, den Dunnen et al. 2012) and ribosomal (Sims, Sudbery et al. 2014) transcript depletion are shown to increase sensitivity of next-generation sequencing expression profiling. It is not evident whether depletion of abundant matrix transcripts would have a comparable effect for microarray-based profiling. Furthermore, by ensuring that a small, enriched cell population was obtained lower RNA concentrations could be used as the reduced heterogeneity in the sample may also reduce the signal-to-noise ratio (Nygaard and Hovig 2008). A critical limitation to use of rat native tendon and fibrin constructs was the number of cells harvested, therefore, matrix depletion could have additional benefit for samples with low cellularity.

3.1.2: Study hypothesis

It was hypothesized that this approach to reduction of sample complexity, by matrix depletion, across all conditions would standardize sample handling with isolated, matrix-depleted cells the resultant input for transcriptomic and proteomic survey. A study consisting of parallel gene and protein surveys would facilitate analysis of prevailing regulatory mechanisms at two levels within the biological hierarchy. A contemporary and well-annotated platform was considered to provide a more robust survey of the transcriptome. By surveying only isolated cells from all conditions it was hypothesised that consistent network perturbations would emerge using pathway topology approaches.

3.1.3: Review: the phosphoinositide-3 kinase (PI-3K)/Akt pathway

The PI-3K pathway is a core signal transduction mechanism for a plethora of physiological cellular mechanisms including differentiation, survival, apoptosis, metabolism and protein synthesis (Cantley 2002). When mutated regulators permit unrestricted activation the PI-3K pathway is responsible for a number of cancers and is also implicated in inflammatory and autoimmune conditions, for example rheumatoid arthritis and systemic lupus erythematosus (SLE) (Foster, Blunt et al. 2012).

PI-3 kinases are a conserved family of lipid kinases; they may be divided into four sub-families, or classes (I-IV), on the basis of their substrate specificity, primary structures, regulation and content of their domains. Class I PI-3 kinases are the best described and consist of four isoforms, α-δ, and are receptor regulated.

The PI-3 kinases may be activated by a number of ligands, including IGF-1, through various receptor tyrosine kinases. Phosphorylation of hydroxyl groups on membrane-bound inositol phospholipids, by PI-3 kinases, allows these to act as secondary messengers. Considering only the receptor-regulated Class I PI-3 kinases these are phosphatidylinositol-4,5,-bisphosphonate (PtdIns(4,5)P2)/PIP_2 kinases that generate PtdIns(3,4,5)P3/PIP_3 following phosphorylation of the 3-OH moiety on inositol membrane lipids. The conversion of PIP_2 to PIP_3 is a key step in the initiation of the signalling cascade (Cantley 2002). The broad range of ligands, secondary messengers, downstream targets and feedback loops makes the PI-3K pathway complex (Vanhaesebroeck, Stephens et al. 2012). A cursory overview of the receptor-regulated Class I PI-3 kinase signal transduction pathway is provided in **Figure 3.1**. Ligand binding to receptor tyrosine kinases results in the auto-phosphorylation of tyrosine residues. PI-3 kinases are recruited to the

membrane. The catalytic p110 subunit is constitutively bound to the p85 regulatory subunit; the latter has an SH2 domain allowing interaction with phosphorylated Tyr (pTyr) and adaptor molecules. This brings the Class I PI-3 kinases in contact with their lipid substrates. Activation by the appropriate tyrosine kinase receptors, for example IGFR1, stimulates the rapid phosphorylation of PIP_2 to PIP_3. These lipid secondary messengers recruit a number of signalling proteins with the pleckstrin-homology (PH) domain to facilitate their activation. Proteins with these domains include the serine/threonine kinase Akt, below. The binding of PIP_2 to PIP_3 drives the translocation of Akt to the plasma membrane where it is phosphorylated at the Thr_{308} and Ser_{473} residues by phosphoinositide-dependent kinase (PDK1) and mammalian target of rapamycin complex 2 (MTORC2) respectively.

The best-described effector of PI-3K signalling is the serine/threonine kinase AK, also known as protein kinase B – more commonly this signaling pathway is written as the PI-3K/Akt signalling pathway. As with all effectors of Class I PI-3 kinases Akt has a pleckstrin homology (PH) domain, which binds directly to phosphorylated inositol lipids, facilitating the Thr_{308} phosphorylation of Akt by the phosphatidylinositide-dependent kinase, Pdk1. More than 100 Akt substrates have been identified, these include: p21, p27, BCL2, FoxOs and GSK-3 (Foster, Blunt et al. 2012). Signalling is inhibited by the action of the phosphatase and tensin homolog (PTEN) by inducing the dephosphorylation of PIP_3 back to PIP_2. The dysregulation of PTEN is a significant contributor to tumourigenesis. Numerous small molecule inhibitors of the PI-3K signalling pathway are described, including non-specific, reversible inhibition by LY294002 (McNamara and Degterev 2011).

Akt is only one example of a number of PI-3K effectors, which, when activated, result in a phosphorylation cascade. Downstream targets regulate a diverse range of physiological functions including: proliferation, cell growth, metabolism, protein synthesis, differentiation and apoptosis.

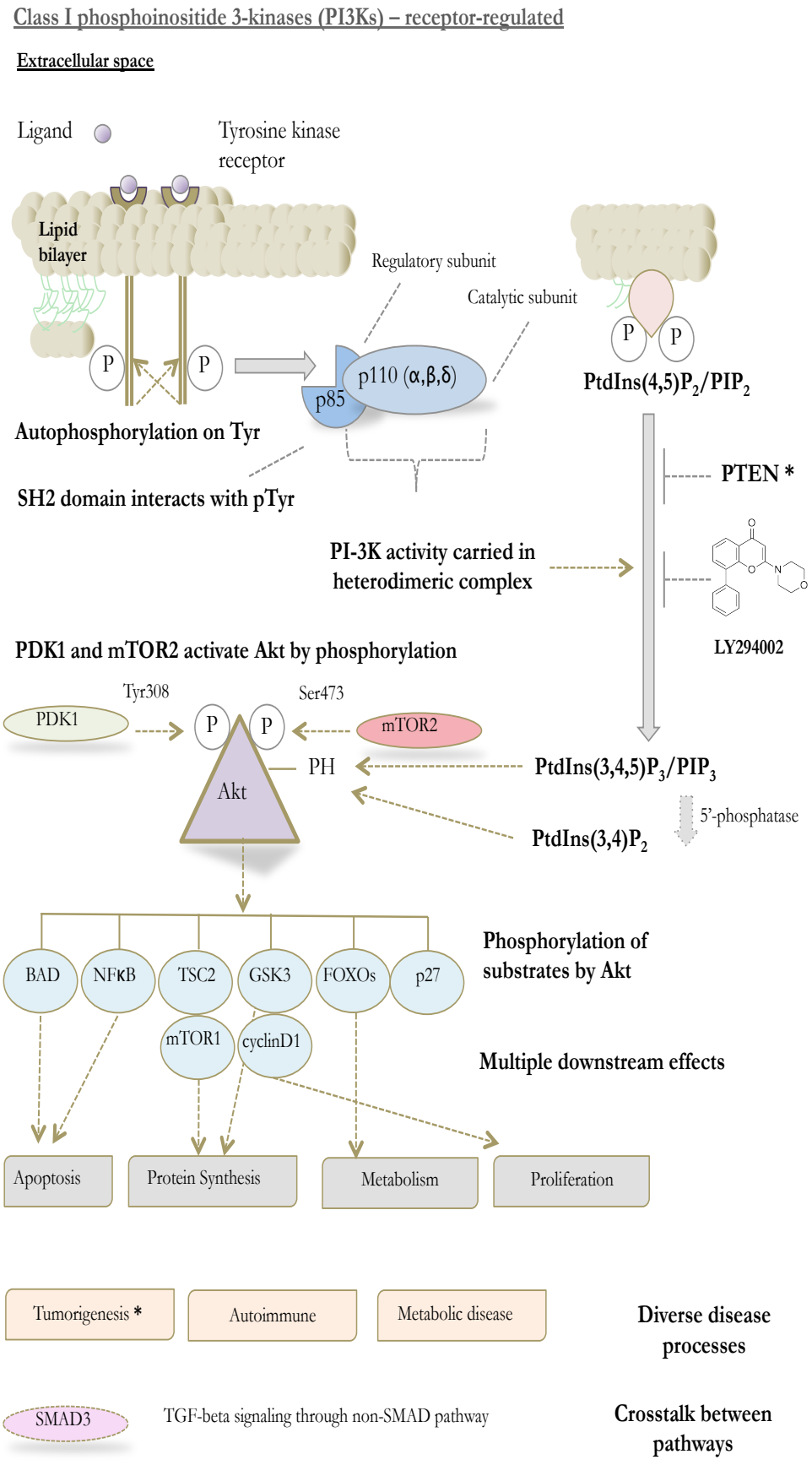


Figure 3.1: Schematic diagram to represent the PI-3K/Akt signaling pathway for class I PI-3 kinases. Full description of the pathway is made in the main text

3.2: Methods

3.2.1: Samples

12 week-old, male, Lewis (n = 12, isogenic, Charles River) rats were obtained (as per **Chapter 2**), **Figure 3.2**. For each biological replicate cartilage was obtained from the coxo-femoral and femoro-tibial joints and pooled (0.63 g ± 0.23, mean and s.d., wet weight); tendon was obtained from the Achilles tendon of the gastrocnemius muscle and the deep digital flexor tendon (0.59 g ± 0.28), these were also pooled. Reagent sources are as described in **Chapter 2** unless otherwise stated.

Biological replicates were divided into three groups (n = 4 in each): Group I - primary tissue-derived cells; Group II - cells that underwent dedifferentiation in two-dimensional culture by serial passage on three occasions; Group III - cells that underwent dedifferentiation in two-dimensional culture then, at the end of the second passage, were transferred to three-dimensional culture systems appropriate for chondrocytes or tenocytes as previously described. The extra- and peri-cellular matrix was depleted in all samples as described below.

Group I: Cells derived from native tissue

Samples were dissected and minced on sterilised, glass petri dishes. Samples were washed twice in DMEM, free of phenol-red and serum, containing only antibiotic and anti-fungal agents. Samples were placed in 50 mL Falcon tubes containing a total volume of 10 mL of 0.4 % collagenase type II prepared with complete media (CM1) and incubated for > 20 hours. Following this, samples were centrifuged: samples from cartilage at 500 x *g* for 8 minutes; samples from tendon at 1000 x *g* to ensure that partially digested fascicular material, containing linear cell arrays in

peri-cellular matrix, would pellet (Ritty, Roth et al. 2003). The supernatant was removed and replaced with 10 mL of 0.25 % trypsin from bovine pancreas (Sigma, #T1426) resuspended in serum-free DMEM. Samples were incubated at 37 °C for four hours. After this period samples were centrifuged as before, washed with complete media, re-suspended in 10 mL and passed through a 70 μm sterile cell strainer to produce a single-cell suspension. A 10 μL sample was obtained and cells counted and evaluated for viability using the trypan-blue exclusion test. Cells for native cells analysis were pelleted and stored in 1 mL of TriReagent® at -80 °C.

Group II: Cells expanded in monolayer culture

Chondrocytes and tenocytes were expanded to passage three. Primary cultures for both tissues started with $2.5x10^5$ cells seeded at $10^4/cm^2$. At confluence cells were split and seeded as for primary culture. Cells in two- and three-dimensional culture were both grown in complete media as described before, but with DMEM with the absence of the phenol-red indicator and with the addition of 200 mM L-glutamine (Gibco, Invitrogen). At passage three cells were dissociated from monolayer, centrifuged and re-suspended in 10 mL 0.4 % (w/v) collagenase type II and incubated at 37 °C for one hour. Cells were washed in PBS twice, counted and live/dead stained.

Group III: Cells retained in three-dimensional cultures

Following monolayer expansion chondrocytes were suspended in alginate following polymerisation in calcium chloride at $4x10^6$ cells mL^{-1} of sterile-filtered alginate. Cells were maintained in 25 mL of complete media 3 (CM3) as described previously with media changed every other day. Cells were released from alginate on day 7. Cell pellets were digested in 0.4 % collagenase as above. Monolayer expanded tenocytes were used to prepare constructs using fibrin gels and cultured

until linear tendon-like constructs were formed. Six technical replicates for each biological replicate (n = 4) were pooled and then digested directly in 0.4 % collagenase type II for 60 minutes. After this period the suspension of cells and fibrillar material was passed through a 70 μm cell strainer. Cell pellets were treated as for Group II monolayer.

RNA preparation and integrity

RNA was extracted and prepared as described in **Chapter 2**. Samples were standardised to a concentration of 20 ng/μL (total volume 25 μL) and all samples had a 260/280 ratio of > 1.8. Samples were stored in Lo-Bind Eppendorf tubes, as before, at -80 °C. Samples were submitted to Molecular Genetic Services, Hologic, Manchester, for microarray analysis. Bioanalyzer analysis provided RIN scores of > 8 for all samples with the majority of samples scoring a maximum of 10 indicating minimal RNA degradation.

3.2.2: Microarray analysis and bioinformatics

Whole transcriptome profiling of 24 samples was undertaken using the GeneChip® Rat Gene 2.0 ST Arrays (Affymetrix, Inc., Santa Clara, USA) with six arrays per chip and all samples prepared in parallel and scanned on the same day. The microarrays interrogated expression of 28,407 RefSeq transcripts. Total RNA was amplified using the Affymetrix GeneChip® WT PLUS Reagent Kit according to manufacturer's instructions. The resulting cDNA was quantified using optical density (NanoDrop, Thermo Scientific). The cDNA was normalised and hybridised onto Affymetrix Mouse Gene ST 2.0 microarrays for 16 hours at 45 °C. Microarrays were washed and stained using the Affymetrix GeneChip® Hybridization, wash, and stain kit were used according to manufacturer's instructions using the Affymetrix GeneChip® Fluidics Station 450. Microarrays

were scanned using an Affymetrix GeneChip® 7G microarray scanner. Data quality control was analysed using Affymetrix® Expression Console™ Software.

Raw .CEL files were imported into *R* (as before) and pre-processed, annotated and statistically assessed as described in the supplied code, **Appendix 3**. Specifically the `oligo` package (Carvalho and Irizarry 2010) was used for pre-processing of .CEL files and the `limma` package for differential expression analysis, as before. All arrays passed quality control thresholds and were retained for further analysis. Filtering, statistical thresholds for inclusion in bioinformatic analysis and methodology (including pathway topology and Ingenuity Pathway Analysis) are consistent with methods presented in **Chapter 2**. Normalised expression data was also used an input to analysis in **Chapters 4** and **7**.

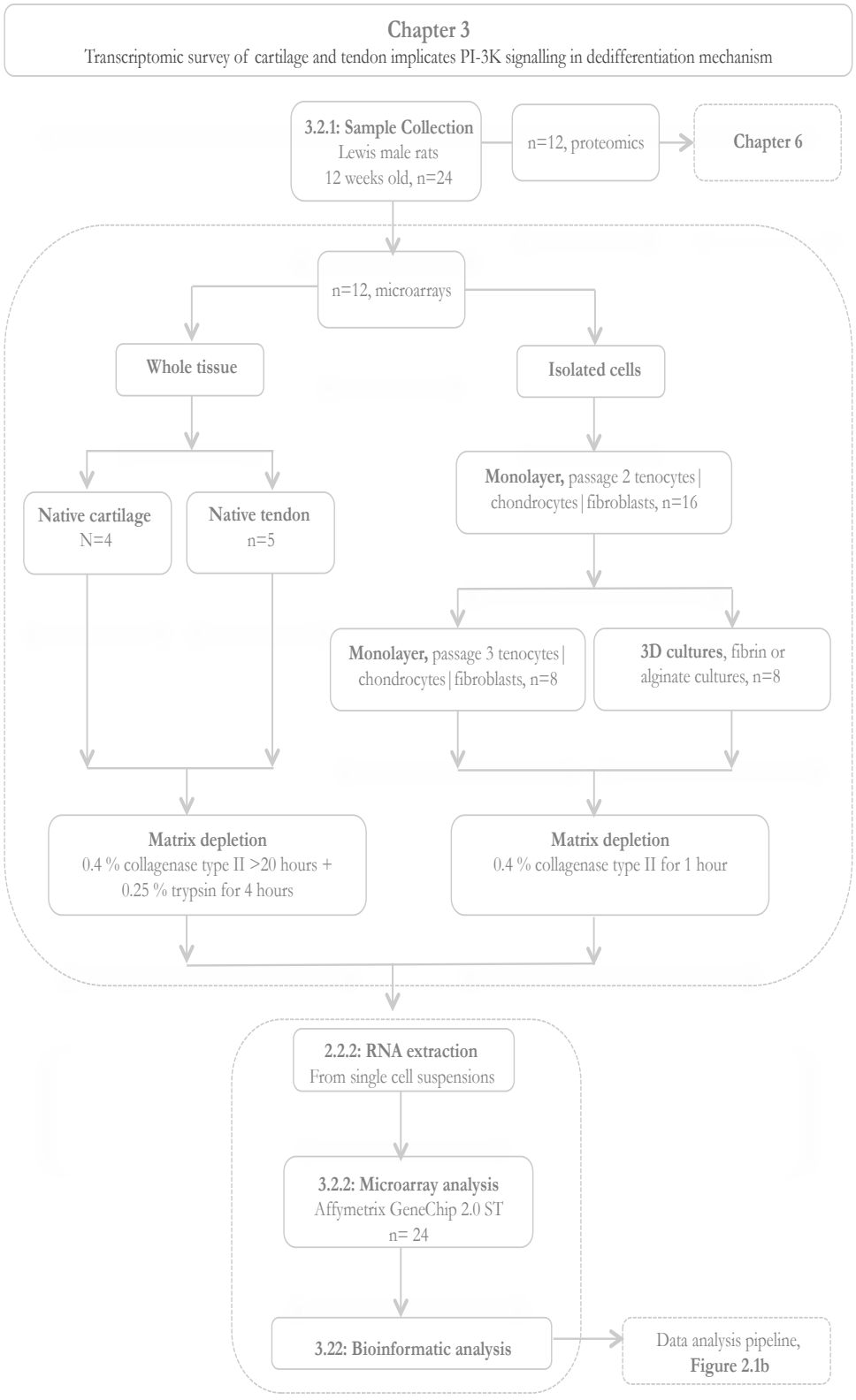


Figure 3.2: Overview of experimental design for results presented in **Chapter 3**. Source of samples for microarray data is indicated (n = 12). Two tissues, cartilage and tendon, were harvested. For each condition (native, monolayer or 3D cultures) there were four biological replicates. Extra- and peri-cellular matrix was depleted by enzymatic digestion as described. Gene expression profile was surveyed using the Affymetrix GeneChip Rat 2.0 ST array, 24 arrays performed in parallel. Data analysis pipeline follows that described in **Figure 2.1b** with microarray platform appropriate *R* packages as described in **3.2.2**. Expression data also used in subsequent analysis in **Chapters 4** and **7**.

3.3: Results

3.3.1: Quality Control

Quality controls of microarrays

Probe-level intensity data was used to produce pseudo-images of the arrays to determine the presence of any systematic or random spatial artifacts across the array, **Figure 3.3**. Fitting of a probe-level model to normalised data did not demonstrate any arrays that significantly departed from expected standards in NUSE and RLE plots, **Figure 3.4**. High correlation was also evident between replicates within a condition, **Figure 3.5**. There was no evidence from quality control analysis that any of the arrays should be removed.

Comparing normalisation techniques

To define whether the cyclic LOESS normalisation strategy, described in **Chapter 2**, was applicable to this dataset it was compared to the standard RMA algorithm (Irizarry, Hobbs et al. 2003), which employs a quantile normalisation method. Background correction methods were identical for both strategies. There was no appreciable difference between the methodologies (slight improvement in the distribution of the densities, **Figure 3.6**, in using a quantile normalisation *versus* LOESS), however, the latter was retained for continuity of analysis between datasets. An uncertainty propagating method, `puma` (Pearson, Liu et al. 2009), was also explored using *R*, but could not be integrated with downstream analysis in subsequent chapters.

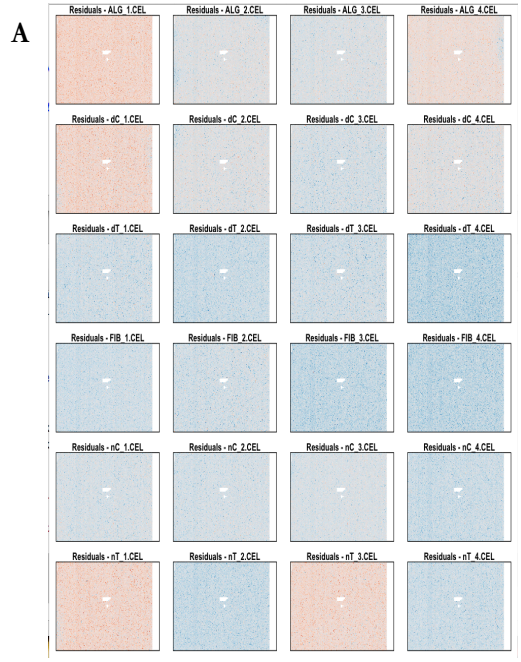


Figure 3.3: Pseudo-coloured images of 24 Affymetrix arrays to visualise spatial artifacts (**A**). The residual images are coloured so that large positive residuals are red, large negative residuals are blue and small residuals are white. Examples of poor quality arrays below (**B**). There do not appear to be systematic technical errors associated with array preparation.

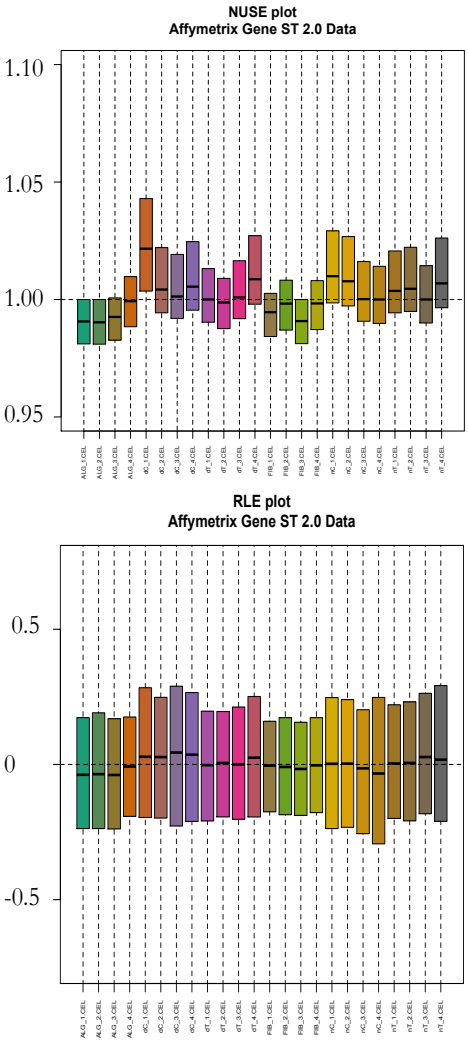


Figure 3.4: The RLE (Relative Log Expression) and NUSE (Normalized Unscaled Standard Error) plots are used assess array quality. Both are derived from a probe-level model (PLM) that computes an expression measure using M-estimator robust regression. **NUSE plot** (top) – shows the normalised standard error (SE) estimates from the PLM such that the median SE=1. Arrays with lower quality are centred higher and have a wider spread. No arrays are centred above 1.1, the threshold for discarding arrays. The **RLE plot** (bottom) uses log-scale estimates for probe expression on each array. For each probe set and array ratios are calculated between a probe set and the median expression value across the same probe set across all other arrays. Relative expression values are presented. Boxes are centered around 0 and have similar range. This is expected given that few genes are differentially expressed across arrays.

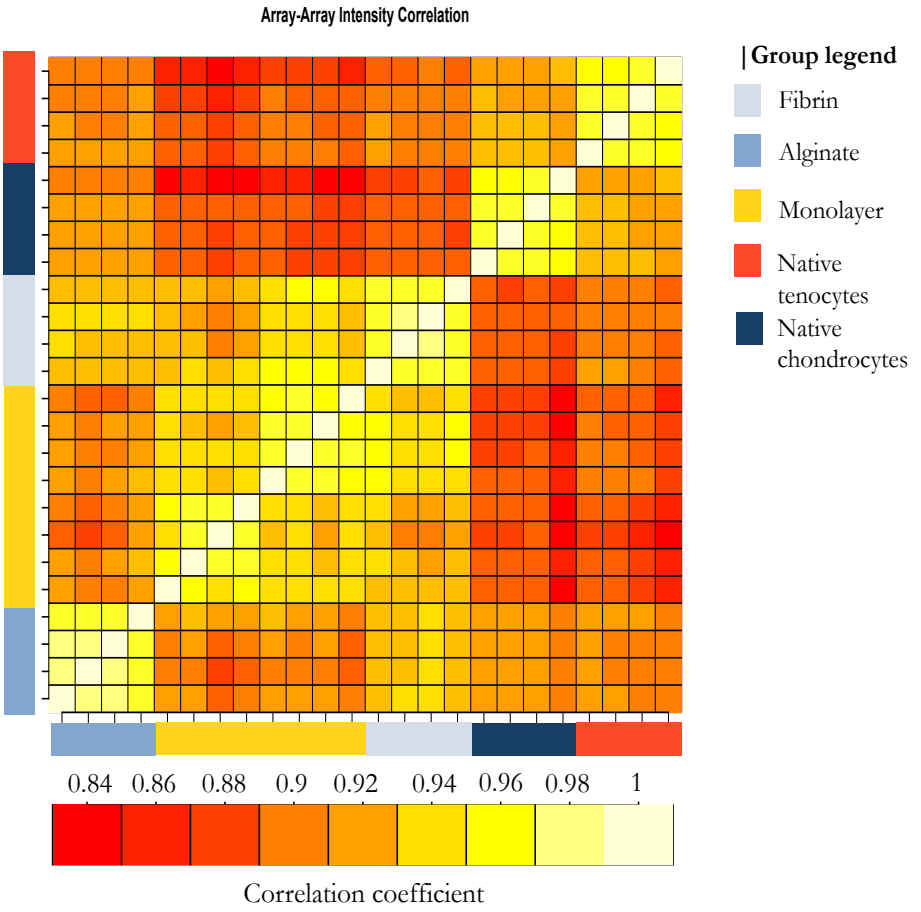


Figure 3.5: Array-array intensity correlations (correlation coefficient) for n=24 Affymetrix arrays profiling chondrocyte and tenocyte expression across different conditions (Group legend).

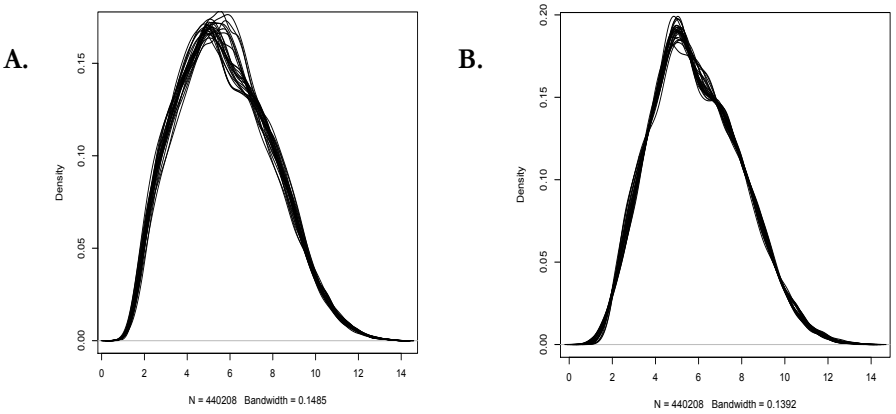


Figure 3.6: Density (y-axis) distribution of normalised, $\log_2$-transformed, intensity data (x-axis) using the RMA (**A**) or LOESS (**B**) methods across 24 arrays. There is moderate improvement in the reproducibility of the sample distribution of the intensities using the LOESS technique for normalisation. Given that this was used for the Illumina data it was elected to continue using this technique for Affymetrix microarray normalisation

3.3.2: Reduction of dimensionality

Principal component analysis

Unsupervised principal component analysis presented four distinct groups defined by experimental condition, **Figure 3.7**. The first two principal components accounted for over 90 % of the variation using expression data from the 500 most co-variant genes (**SD3.1**). Concurring with findings in **Chapter 2** chondrocytes and tenocytes in monolayer culture had convergent gene expression profiles at passage three. Cells derived from whole cartilage and tendon tissue, when matrix-depleted, clustered more closely with each other, whilst cells in either alginate or fibrin cultures were the most divergent and did not cluster together as previously shown. Of the 500 most covariant genes only 31 matched across both Affymetrix and Illumina data sets including, clusterin (*Clu*), transgelin (*Tagln*), *Tgfβ3*, paired-related homeobox 2 (*Prrx2*), integrin α11 (*Itga11*) and glypican 4 (*Gpc4*).

Hierarchical clustering

Unsupervised, hierarchical clustering demonstrated two principal clades consisting of samples from monolayer cultures and fibrin cultures containing tenocytes, and a second consisting of cells isolated from native tissue and chondrocytes in alginate beads, **Figure 3.8**.

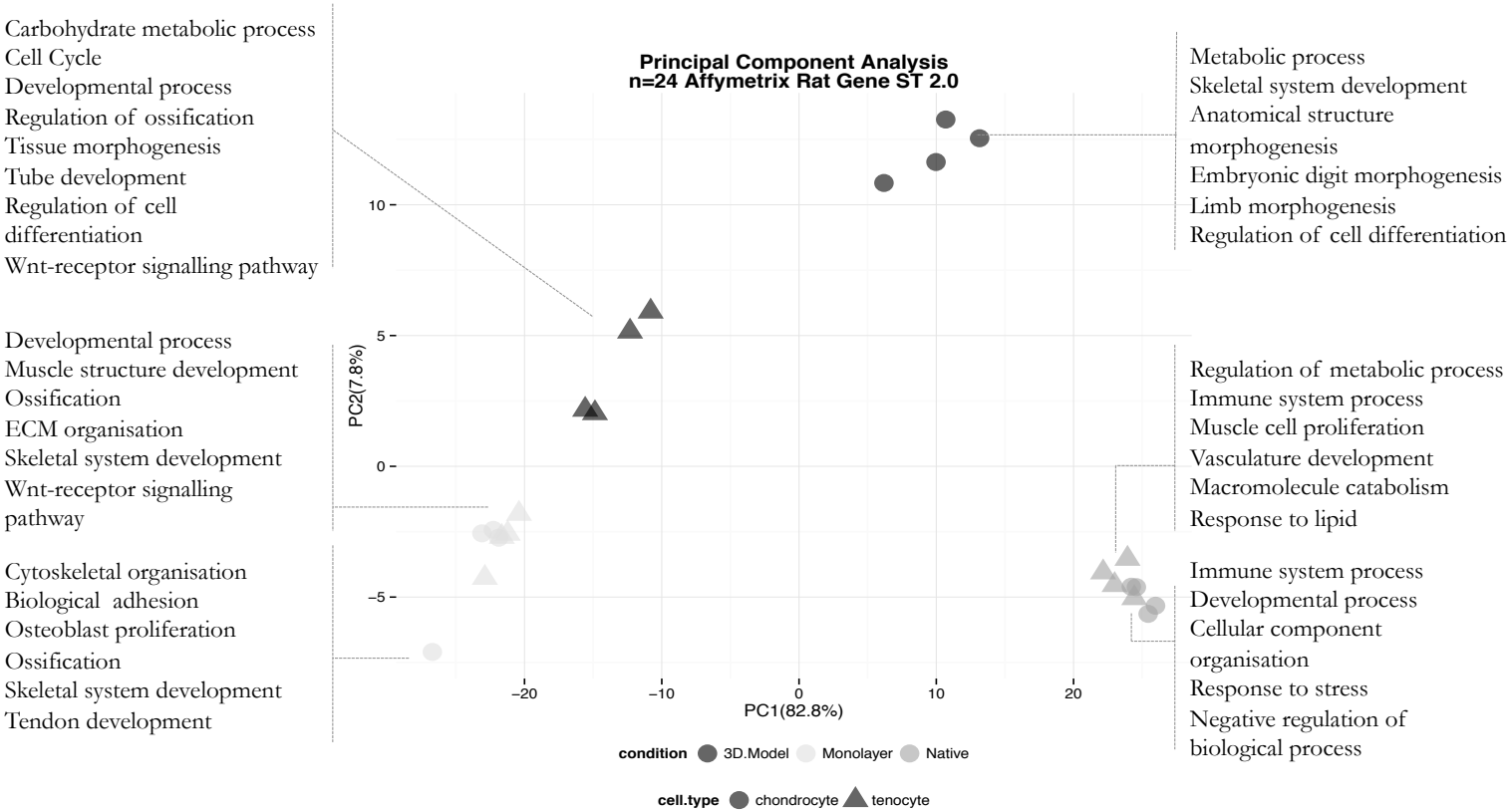


Figure 3.7: Principal Component Analysis. Top 500 most co-variant genes from normalised and filtered expression data from 24 samples (data points indicate individual arrays with sample type defined by the figure key). Chondrocytes and tenocytes isolated from native tissue and digested free of extra-cellular matrix cluster more closely with each other than with monolayer or three-dimensional cultures. Chondrocytes derived from alginate cultures cluster remote from both monolayer and fibrin constructs, but also from native chondrocytes. Gene ontology analysis using a hypergeometric distribution with Entrez annotated probes from Affymetrix Gene ST 2.0 used as background. Redundant terms filtered out using SimRel (ReviGO). Biological Process (BP) ontology annotations are based upon genes more highly expressed in each specific condition relative to the native tissue. Gene ontology analysis was undertaken on all genes passing differential expression thresholds and not restricted to the 500 most covariant genes.

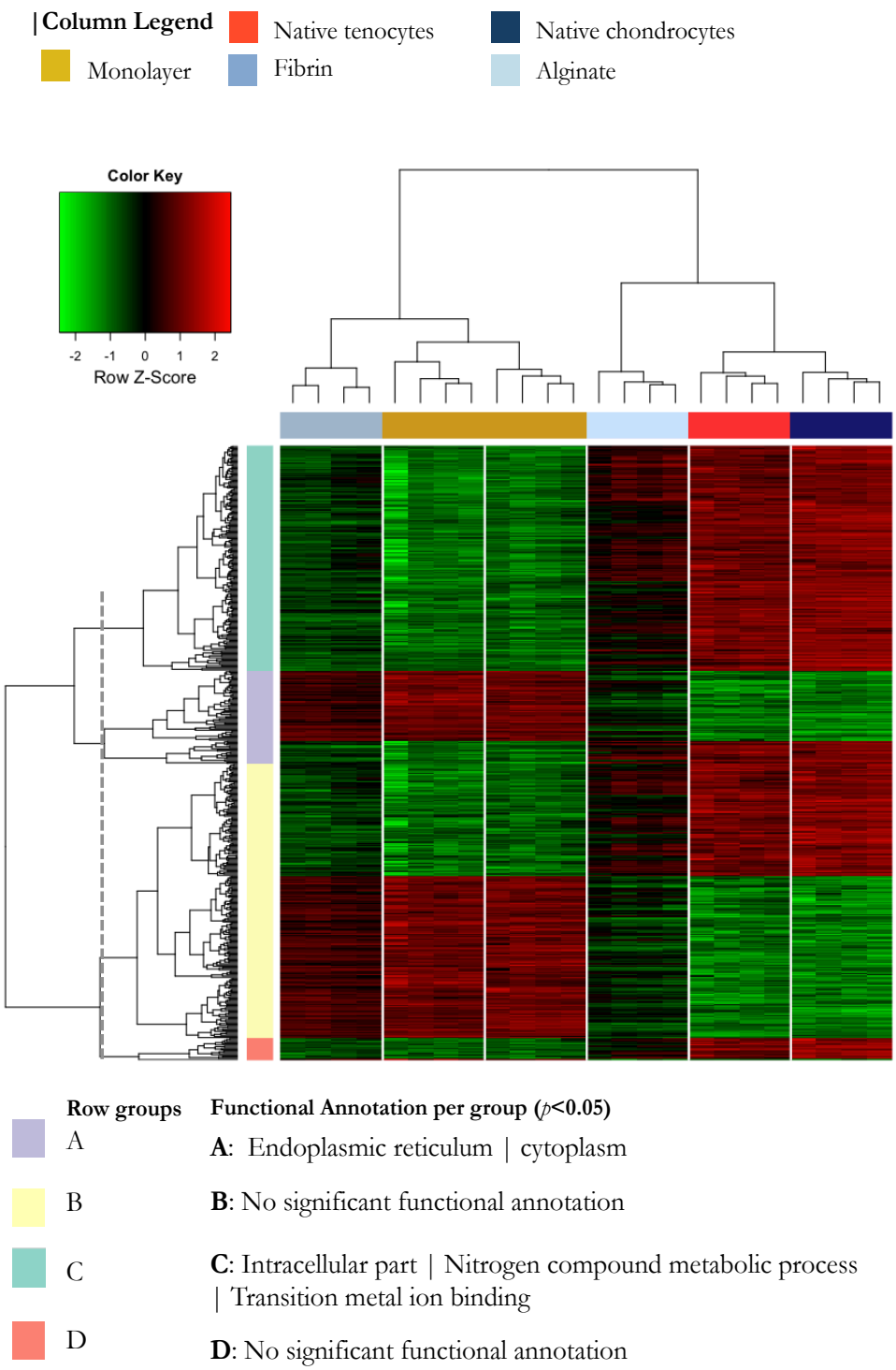


Figure 3.8: Heatmap and hierarchical clustering
Data derived from the 500 most co-variant genes (**rows**) across 24 samples (**columns**). Using unsupervised clustering the data clusters into two main clades, the first defined by monolayer culture and fibrin constructs, the second by isolated native cells and alginate cultures. Expression levels are normalised and scaled by row with red indicating high expression and green indicating low expression as defined by the key. The row groups legend indicates the group associations defined by the dashed line bisecting row clades. Functional annotation, using DAVID, is not significant for some groups. Row annotations and groups in **SD3.1**.

3.3.3: Differential gene expression analysis

Differential expression analysis of arrays reflected those findings from the Illumina dataset, **Chapter 2**. The greatest number of differentially expressed genes was found between cells isolated from native tissue when compared to passage three dedifferentiated chondrocytes or tenocytes – 3863 genes were differentially expressed between native chondrocytes and dedifferentiated chondrocytes. The fewest differentially expressed genes were found between passage three chondrocytes and tenocytes. Notably, the comparison of native chondrocytes to tenocytes presented fewer differentially expressed genes (n=1771) than the comparison between fibrin constructs and alginate beads (n=2251). Full pairwise comparison lists are found in supplementary data **SD3.2-3.10**.

Chondrocytes: Dedifferentiation transition

A summary of the most differentially expressed genes is provided in **Figure 3.9**. At passage three, chondrocytes in monolayer were characterized by the high expression of the actin-binding protein transgelin, *Tagln*, an inhibitor of receptor tyrosine kinase signaling, *Grb14*, and galectin, a beta-galactosidase binding lectin (*Lgals1*). Genes associated with oxidative phosphorylation (ATPase synthase subunits, NADH dehydrogenase complexes) were all more highly expressed in monolayer cells. Additionally, as expected, many genes were associated with cell cycle and microtubule processes. Relative to native chondrocytes, collagen I, III, V, VIII and XVIII alpha subunits were all more highly expressed in monolayer.

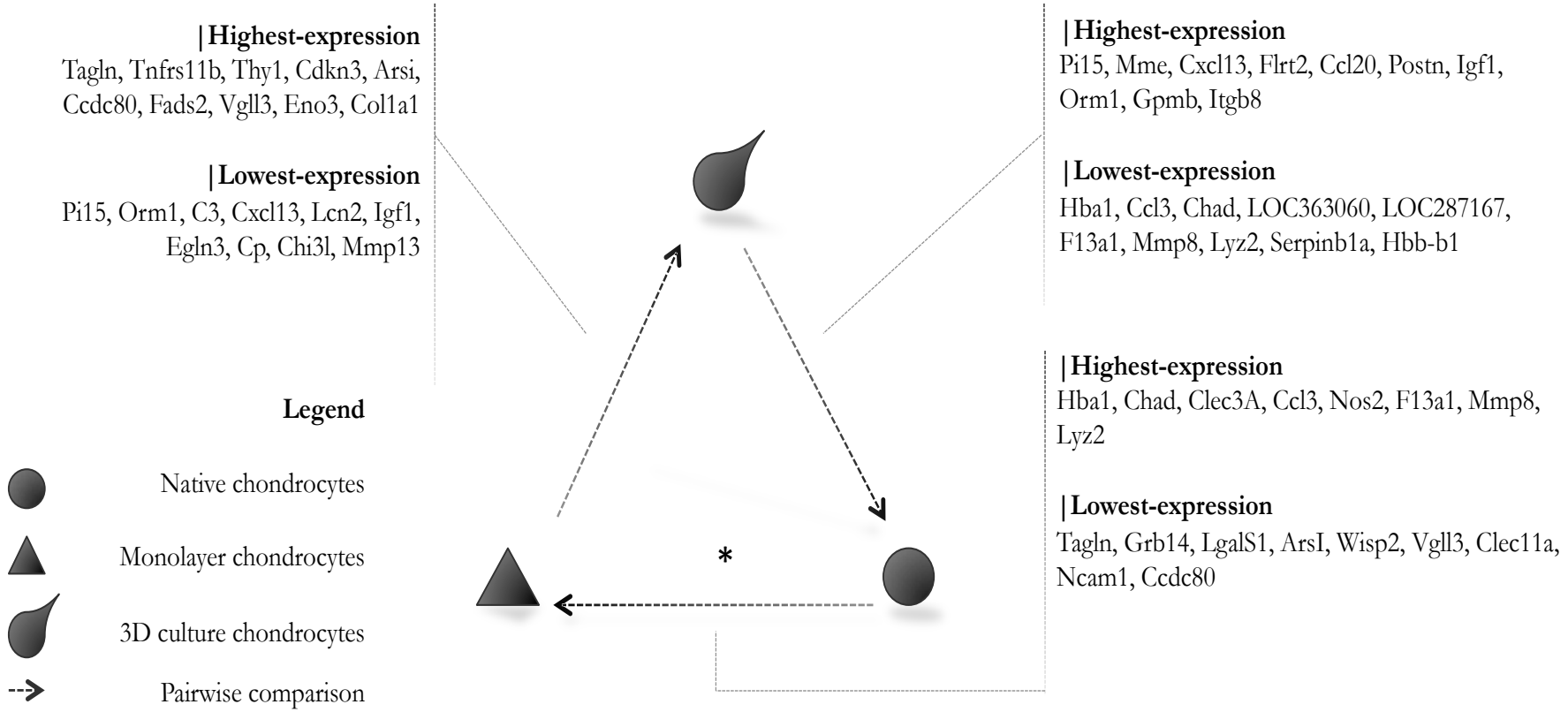


Figure 3.9: Schematic diagram: Genes showing the most extreme differential expression in pairwise comparisons between different environmental conditions for **chondrocytes**. In each case extra- and peri-cellular matrix has been depleted. The direction of the arrow indicates the direction of the pairwise comparison, e.g. in ***** genes are expressed more highly in native chondrocytes relative to the monolayer environment. Full differential expression lists, including full gene annotations, are found in **Supplementary Data 3.2-3.4.** Elements of schematic are defined in figure legend. Genes are not italicised for this figure.

'System development' annotated genes including receptor-Smad *Smad3*, and inhibitor-Smads *Smad6* and *Smad7*, SoxC group genes *Sox12* and *Sox4*, and *Bmp3* (osteogenin) and *Bmp4* were all also more highly expressed in monolayer chondrocytes. Genes with musculoskeletal system development associations Mustang/*Mustn1*, scleraxis/*Scx* and mohawk/*Mkx*, investigated in the qPCR analysis in **Chapter 2**, were more highly expressed in monolayer chondrocytes. The hindlimb specific homeobox gene, *Pitx1*, and *Pitx2* were also up-regulated in this comparison. Members of the transforming growth factor superfamily *Tgfβ1-3* were all represented in this analysis. The presence of non-muscle myosins and tenascin was also confirmed in this study, in addition to the high expression of chemokines and their receptors in cells derived from native cartilage. The CCN family was represented by the higher expression of *Wisp1* and *Wisp2*; mesenchyme-associated *Thy-1* and *Snai1*, presented in **Chapter 2**, were also more highly expressed in monolayer.

In native chondrocytes there was higher expression of type II collagen alpha-1 subunit (*Col2a1*), decorin (*Dcn*), aggrecan (*Acan*), lubricin (*Prg4*), chitinase 3-like 1 (*Chi3l1*), and the matrix metalloproteinases *Mmp3* and *Mmp13*. The bone morphogenetic proteins *Bmp6* and *Bmp2*, interleukin 1-beta (*Il-1b*), and chemokines *Cxcl2* and *Cxcl16* were also more highly expressed in native chondrocytes relative to monolayer.

Chondrocytes: Re-differentiation transition

In alginate beads chondrocytes were characterized by the expression of genes associated with the regulation of chondrogenesis including *Sox4* and SoxD group genes *Sox6* and *Sox5*, differentiation factors *Egr1* and *Egr2,* and *Bmpr1a* and *Bmpr1b*. Numerous chemokines, some implicated in cartilage pathology, were also

more highly expressed in alginate beads relative to monolayer, including: *Pi15*, *Cxcl2*, *Cxcl12, Ccl20*, and interleukins *Il-6* and *Il-7*. The *Col2a1* expression regulators *Sp1* and *Sp3* transcription factors were up-regulated in alginate. Aggrecanases *Adamts5* and *Adamts1* were also more highly expressed. Chondrogenesis and matrix regulators such as *Grem1*, *Frzb*, and *Dkk* were more highly expressed in alginate beads than monolayer or native chondrocytes. Alginate cultures in this study were characterized by the increased expression of *Mmp13*, *Hif1a*, the PI-3K inhibitor *Pten*, and lower expression of *Runx2* relative to monolayer, native cartilage and fibrin cultures.

Monolayer tenocytes

A summary of the most differentially expressed genes in tenocytes is provided in **Figure 3.10**. In addition to *Tagln*, monolayer tenocytes at passage three had higher expression of integrin alpha-11 (*Itga11*), sushi-repeats containing protein *Srpx*, and coiled-coil domain containing protein *Ccdc80*. As shown for monolayer chondrocytes CCN-family members *Wisp1* and *Wisp2* and tendon development-associated genes, *Scx*, *Mkx* and *Mustn1* were all more highly expressed in monolayer tenocytes. Genes associated with SLIT-ROBO neuronal guidance and tendon development in Drosophila, *Slit3* and *Robo2*, were both highly expressed in monolayer tenocytes. Concurring with monolayer chondrocyte expression profiles inhibitor Smads, *Smad6* and *Smad7*, were more highly expressed in monolayer tenocytes than tenocytes from native tendon.

Native tenocytes

In comparison to monolayer tenocytes there was higher expression of tenomodulin (*Tnmd*), clusterin (*Clu*), chondroadherin (*Chad*), lubricin, and chemokines *Cxcl2* and *Cxcl13* in native tenocytes. The expression of superoxide

dismutase (*Sod2*), frizzled B (*Frzb*), SRY (sex determining region Y)-box 5 (*Sox5*), angiopoietin-like 4 (*Angptl4*), was higher in native tenocytes relative to monolayer concurring with the equivalent comparison in native chondrocytes.

Fibrin constructs

Chemokines, including, *Cxcl13*, *Cxcl5* and *Cxcl1* and the interleukins *Il-6*, *Il-1a*, *Il-11* and *Il-33* were all more highly expressed in tenocytes in fibrin constructs. Modulators of the Wnt-signalling cascade, secreted frizzled related proteins *Sfrp1*, *2* and *4*, and the BMP antagonist gremlin 1, *Grem1*, were all more highly expressed in fibrin constructs. Dedifferentiated tenocytes had significantly higher expression of transcripts to *Bmp1*, *3*, *4*, and *6*; in contrast higher expression of *Bmp2* was evident in fibrin constructs. Relative to native tenocytes expression of tenascin N/W, *Tnn*, and the matricellular integrin ligand periostin, *Postn*, were significantly higher in fibrin constructs. Tendon development-associated scleraxis (*Scx*) was also more highly expressed.

Relative to alginate cultures the higher expression of microfibril-associated genes *Mfap4* and *Mfap5*, *Tnn*, *Thbs4*, *Fbln2*, and development regulators *Mustn1* and *Meox2* was demonstrated – these have been presented and discussed in **Chapter 2**. *Runx2* was more highly expressed in fibrin and monolayer cultures relative to native tendon and alginate cultures.

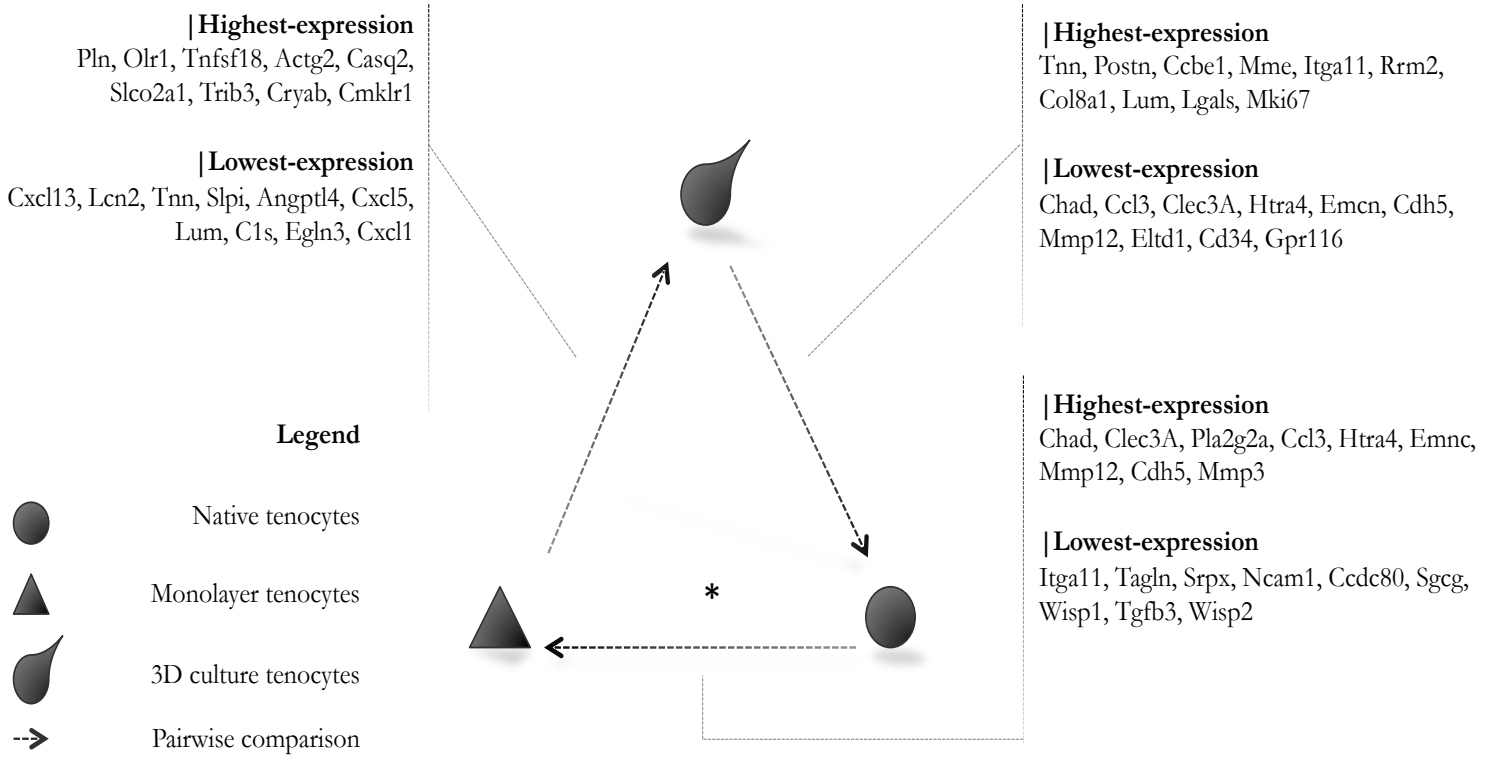


Figure 3.10: Schematic diagram: Genes showing the most extreme differential expression in pairwise comparisons between different environmental conditions for **tenocytes.** In each case extra- and peri-cellular matrix has been depleted. The direction of the arrow indicates the direction of the pairwise comparison, e.g. in ***** genes are expressed more highly in native tenoocytes relative to the monolayer environment. Full differential expression lists, including full gene annotations, are found in **Supplementary Data 3.5-3.7.** Elements of schematic are defined in figure legend. Genes are not italicised for this figure.

3.3.4 Functional annotations

For gene ontology analysis complete, filtered, differential expression lists were used. A summary of significant terms is presented in **Figure 3.7**.

Chondrocytes

Cells isolated from cartilage were annotated with biological process functions: 'immune system process', 'developmental process', 'cellular component organisation', 'response to stress' and 'negative regulation of biological process'.

In comparison, chondrocytes from monolayer expression profiles were described by biological process terms including: 'cytoskeletal organisation', 'biological adhesion', 'developmental process', and 'ossification'. Notably terms such as 'regulation of ossification', 'fibril organisation', 'skeletal system development' and 'tendon development' were all gene ontology terms significantly enriched in this expression profile. Notably, the latter term was only found in monolayer chondrocyte expression data.

Functional annotation of gene transcripts found to be more highly expressed in chondrocytes within alginate beads, relative to their native counterparts, was significantly enriched with terms relating to: 'metabolic process', 'skeletal system development', 'anatomical structure morphogenesis', 'embryonic limb morphogenesis' and 'regulation of cell differentiation'.

Full gene ontology lists are found in **SD3.15**. Using Ingenuity Pathway Analysis general disease, metabolic and physiological functional annotations for chondrocytes are provided in **Table 3.1**.

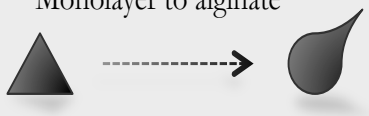
Comparison	IPA descriptors
Native cartilage to monolayer 	**DD**: Cancer| Cardiovascular disease| Gastrointestinal disease; **MCF:** Cell growth and proliferation | Cell movement | Cell death and survival **PS**: Cardiovascular system development and function | Organismal development | Immune cell trafficking | Tissue development
Monolayer to alginate 	**DD**: Cancer | Organismal injuries and Abnormalities |Reproductive system disease **MCF**: Cell growth and proliferation | Cell death and survival | Cellular movement **PS**: Cardiovascular system development and function | Organismal development | Tissue development | Skeletal and muscular system development and function
Alginate to native cartilage 	**DD**: Inflammatory response | Cardiovascular disease | Connective tissue disease | Skeletal and muscular disorders **MCF**: Cellular movement | Cellular growth and proliferation | Cell death and survival **PS**: Immune cell trafficking | Hematological system development and function | Cardiovascular system development and function | Organismal development

Table 3.1: Functional descriptors of differentially expressed gene lists derived using Ingenuity Pathway Analysis. Legend: **DD** – Diseases and Disorders; **MCF**: Molecular and Cellular Function; **PS**: Physiological System Development and Function. Within each parent summary term lie numerous 'child' terms carrying their own *p*-values and activation *z*-scores. In the comparison 'alginate to native cartilage' the term 'Skeletal and muscular disorders' defines 'Rheumatic Disease' as the top child term (p=4.31e-24) with a predicted decreased activation (z=-2.99).

Comparison	IPA descriptors
Native tendon to monolayer 	**DD**: Cancer | Cardiovascular disease | Inflammatory response **MCF:** Cell growth and proliferation | Cell death and survival | Cellular movement **PS**: Cardiovascular system development and function | Tissue morphology |
Monolayer to fibrin 	**DD**: Inflammatory response | Cancer | Organismal injury and abnormalities | Cardiovascular disease **MCF**: Cellular movement | Cellular growth and proliferation | Cellular development **PS**: Cardiovascular system development and function | Organismal development | Embryonic development | Immune cell trafficking
Fibrin to Native tendon 	**DD**: Cancer| Cardiovascular disease | Organismal injury and abnormalities **MCF**: Cellular movement | Cellular growth and proliferation | Cell death and survival **PS**: Cardiovascular system development and function | Organismal development | Organismal survival | Tissue development

Table 3.2: Functional descriptors of differentially expressed gene lists derived using Ingenuity Pathway Analysis. Legend: **DD** – Diseases and Disorders; **MCF**: Molecular and Cellular Function; **PS**: Physiological System Development and Function. Terms for tenocyte comparisons show consensus with the same comparisons between chondrocyte conditions.

Tenocytes

In tenocytes isolated from native tissue gene ontology biological process annotations included: 'immune system process', 'regulation of metabolic process', 'muscle cell proliferation, 'vasculature development', and 'macromolecule catabolism', **SD3.16**.

In comparison, monolayer tenocytes at passage three were annotated with biological process terms including: 'developmental process', 'muscle structure development', 'ossification', 'extracellular matrix organisation', 'skeletal system development' and 'Wnt-receptor signaling pathway'.

Those isolated from fibrin constructs were represented by gene ontology terms including: 'carbohydrate metabolic process', 'cell cycle', 'developmental process', 'regulation of ossification', 'tissue morphogenesis', 'regulation of cell differentiation' and 'tube development'.

IPA descriptors for tenocytes are presented in **Table 3.2**. Functional terms associated with cardiovascular disease and cardiovascular system development and function were common to both chondrocyte and tenocyte analyses.

In comparisons of three-dimensional culture systems alginate beads were significantly enriched for general terms associated with, 'metabolic process', 'organic substance metabolic process' and 'nitrogen compound metabolic process', but additionally with more specific terms, including – 'stem cell differentiation' and 'anatomical structure formation involved in morphogenesis'. In contrast, cells isolated from fibrin cultures had a gene expression profile that was reflected in the terms: 'cytoskeletal organisation', 'cardiovascular system development', 'cell adhesion', 'regulation of anatomical structure morphogenesis', **SD3.17**.

3.3.5: Consensus of differentially expressed genes across cell types

As shown in **Figure 3.3** the depletion of peri-cellular matrix resulted in a co-clustering of native tenocytes and chondrocytes relative to other conditions. A pairwise comparison between native tenocytes and chondrocytes and monolayer equivalents was prepared with the intention of defining a consensus gene expression profile for dedifferentiation, **SD3.11**. There were 2538 genes that were common to the dedifferentiation transition for chondrocytes and tenocytes. In **Figure 3.11** an overview of consensus gene expression and tissue-specific findings are presented.

Functional annotation of native to monolayer dedifferentiation genes

Gene ontology biological process functional annotations (DAVID) were used to present consensus genes in a functional context; the number of genes and false discovery rate (FDR) for each term are provided. Genes listed were contained within each gene ontology term and were common to both chondrocytes and tenocytes.

Native chondrocytes and tenocytes shared genes with functional annotations relating to 'regulation of cellular process' (FDR=1.6e-26; 444 genes; *Sp1*, *Sp3*, *Sp4*, *Timp3*, *Atf3*, *Igf2*), 'regulation of gene expression' (FDR=3.6e-13, 197 genes; *Klf4*, *Klf6* and *Klf9*, *Pou2f1*, *ApoE*, *Foxo1*, *Foxo3*), 'response to oxygen levels' (FDR=8.7e-4; 30 genes; *Hif1a*, *Tlr2*, *Tgfb1*, *Angptl4*), 'immune response' (FDR=5.7e-5; 53 genes; *Ccl3*, *Il-1a*, *Il-1b*), 'regulation of cell differentiation' (FDR=2.3e-4; 59 genes; *Sox5*, *Klf4*, *Bmp2*, *Clu*, *Fgf9*, *Tgif1*, *Jun*).

Dedifferentiated chondrocytes and tenocytes in monolayer shared functional annotations related to: 'cell cycle' (FDR=2.1e-6; 68 genes; *Ilk*, *Ccnb2*, *Aurkb*);

'cytoskeleton organisation' (FDR=3.2e-6; 52 genes; *Dbn1*, *Eln*, *Itgb1*, *Myh10*, *Thy-1*); 'extracellular matrix organisation' (FDR=1.3e-3; 20 genes; *Adamts2*, *Ccdc80*, *Col1a1*, *Col3a1*, *Tgfb2*) and 'developmental process' (FDR=3.4e-5; 237 genes; *Meis1*, *Smad3*, *Smad6*, *Smad7*, *Fgf10*, *Fzd1*, *Fzd2*, *Meox2*, *Mkx*, *Pitx1*, *Snai1*).

Genes unique to a tissue comparison were extracted from the expression profiles and are available in **SD3.12-3.13**.

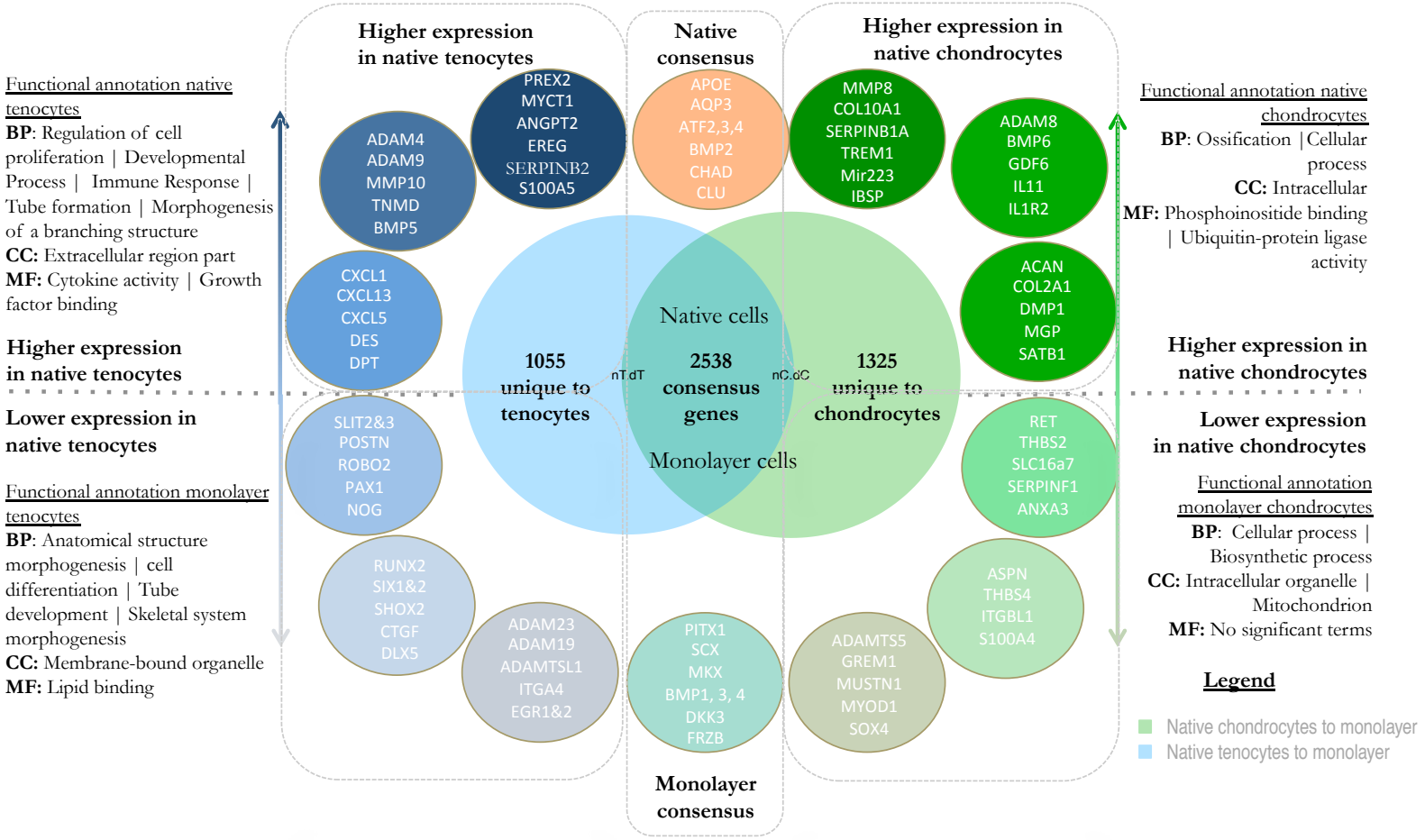


Figure 3.11: Euler diagram: Overlap between differentially expressed gene lists was evident in Affymetrix data in the native cell to monolayer transition. There were 2538 genes that were differentially expressed in common. Genes that were unique to a tissue-specific comparison were considered in terms of their $\log_2$ fold change and annotated with gene ontology terms. Selected genes are provided based upon the fold-change and relevance in the literature. Gene ontology terms were all significant at $p<0.05$ after Benjamini-Hochberg (FDR) correction.

3.3.6: Comparison of independent data sets identifies tissue-associated genes targets for expression validation

In order to explore gene expression profiles that may be unique to native cartilage or native tendon, and so isolate tissue-associated gene profiles within heterogenous samples, differentially expressed genes were considered from both Illumina and Affymetrix data sets. Only those genes differentially expressed in both studies were considered. Across the two datasets there were 311 genes common to both differential expression lists, of which 71 did not match in the direction of the fold-change, **Figure 3.12, SD3.14**. There was a moderately high correlation between the fold changes of these common genes (non-matches included), cor = 0.66, p < 2.2e-16).

For tendon-derived cells several genes were identified which showed higher expression relative to cartilage, with $\log_2$ fold change > 1.5 in both data sets, including: *Tnmd*, *Serpinf1*, *Igfbp6*, *Cxcl13*, *Cpxm2*, *Mfap5*, and *Aspn*. Within the gene lists *Meox2*, *Mustn1*, *Thbs4* and *Thbs2*, and *Prrx1* were also represented, but had more divergent fold changes between data sets. When these 'tendon-associated' genes were functionally annotated using gene ontology terms the following biological process and cellular compartment terms were significant (p<0.05): 'blood vessel development', 'developmental process', 'extracellular matrix'. Tendon or muscle-associated gene ontology terms were not present.

For cartilage-derived cells considered relative to tendon high consensus expression was noted for *Col2a1*, *Mmp8*, *Serpinb1a*, *Sell* and *Ibsp*. Cartilage-associated genes were annotated with the following functional terms: 'defense response', 'extracellular region', 'bio-mineral formation'.

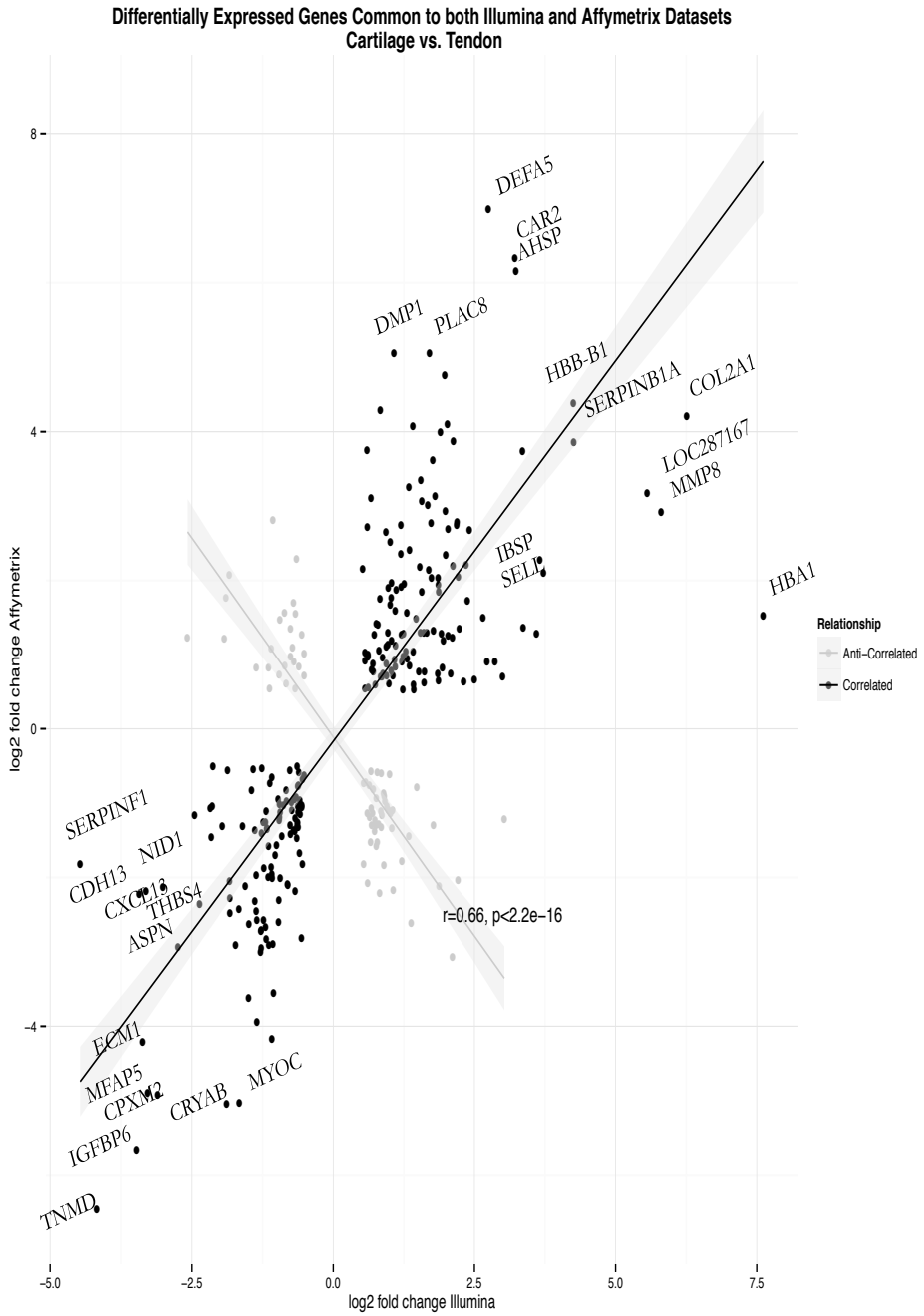


Figure 3.12: Correlation scatterplot:

Genes found to be differentially expressed between native cartilage (top right quadrant) and tendon (bottom left quadrant) in both Illumina (x-axis) and Affymetrix (y-axis) dat sets plotted by $\log_2$ fold change. For clarity only some data points are annotated. Full lists are available in **SD3.14**. Data points are defined in figure legend. Where genes have the same directional change the data point is defined as a dark dot, whereas genes with conflicting (anti-correlated) expression changes are defined as grey points (see relationship legend). For **tendon** *Tnmd*, *Igfbp6*, *Serpinf1*, *Mfap5* and *Ecm1* are all highly expressed in two independent datasets. For **cartilage** *Col2a1*, *Serpinb1a*, *Mmp8*, *Defa5* and *Ibsp* are confirmed to be highly expressed.

3.3.7: Pathway topology analysis defines PI-3K signalling as differentially activated between monolayer and three-dimensional cultures

The activity status of canonical KEGG pathways was inferred by SPIA using all filtered genes within a pairwise comparison and their $\log_2$ fold change as the effect size. Pathways significant after FDR adjustment are presented. The conditions are represented in **Figures 3.13-3.14**.

Chondrocytes

The transition from native chondrocytes (baseline condition) to the dedifferentiated state was defined by activation of the canonical KEGG pathways: 'osteoclast differentiation', 'chemokine signalling pathway' and 'PI-3K/Akt pathway'. With monolayer chondrocytes as the baseline the re-differentiation transition was defined by activated pathways: 'focal adhesion', 'cell cycle' and 'extra-cellular matrix-receptor interaction'; the 'PI-3K/Akt pathway' was predicted as inhibited, along with the associated 'HIF-signalling' and 'FOXO-signalling' pathways. Comparing alginate beads to native chondrocytes the most significant, activated pathways were: 'PI-3K/Akt pathway', 'cytokine-cytokine receptor interaction' and 'Ras signalling pathway'. Full lists are available in **SD3.18**.

|IPA
- **Hepatic stellate cell activation**;
- Acute phase response signalling
- Axonal guidance signalling.
- Role of osteoblasts, osteoclasts and chondrocytes in rheumatoid arthritis

|KEGG

+ Focal adhesion;
+ Cell Cycle;
+ ECM-receptor interaction.

- HIF-1 signalling pathway;
- FoxO signalling pathway;
- **PI3K-AKT signalling pathway.**

Legend

Native chondrocytes

Monolayer chondrocytes

3D culture chondrocytes

Pairwise comparison

Common upstream regulators

TP53, TGFβ1, HRAS, HGF, PDGF BB, TNF, IL6, IL1B

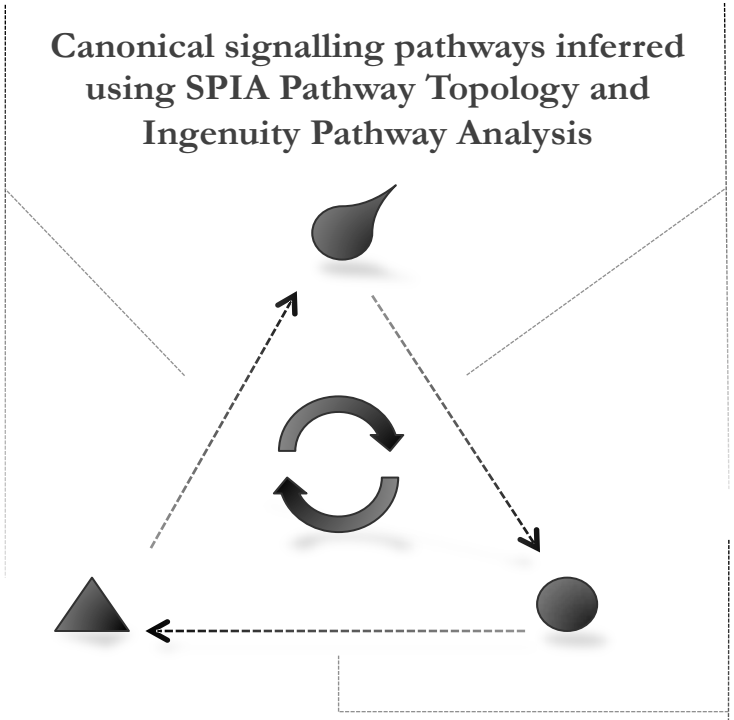


Figure 3.13: Schematic diagram. Differentially expressed genes from chondrocyte comparisons used for pathway topology analysis and pathway prediction. Pathways are predicted as activated (+) or inhibited (-). Arrows indicated the direction of the comparison. Most significantly perturbed KEGG pathways are shown; Ingenuity (IPA) canonical pathways enriched with the same data sets are shown. Common upstream regulators are predicted by IPA (box).

|IPA
- Granulocyte adhesion & diapedesis;
- Atherosclerosis signaling
- LXR/RXR activation;
- **Hepatic stellate cell activation.**

|KEGG

+ **PI3K-AKT signalling**
+ Cytokine-cytokine receptor interaction
+ Ras-signalling

- ECM-receptor interaction
- Focal adhesion
- Osteoclast differentiation

|IPA
- **Hepatic stellate cell activation**;
- Atherosclerosis signaling
- Aryl hydrocarbon receptor signalling;
- p53 signalling;
- Integrin signalling.

|KEGG

+ Osteoclast differentiation
+ Chemokine signalling pathway
+ **PI3K-AKT signalling**

- Focal adhesion
- Cell cycle
- FoxO signalling pathway

|IPA

- Role of osteoblasts, osteoclasts, and chondrocytes in rheumatoid arthritis,
- **Hepatic stellate cell activation**,
- Granulocyte adhesion and diapedesis
- Role of macrophages, fibroblasts and endothelial cells in rheumatoid arthritis

|KEGG

+ MAPK signalling pathway
+ Transcriptional mis-regulation in cancer
+ ECM-receptor interaction

- Cytokine-cytokine receptor interaction
- HIF-1 signalling pathway
- **PI3K-Akt signalling pathway**

Legend

Native tenocytes

Monolayer tenocytes

3D culture tenocytes

Pairwise comparison

Common upstream regulators

TP53, TGFβ1, PDGF BB, TNF, IL1B, LPS, ERBB2

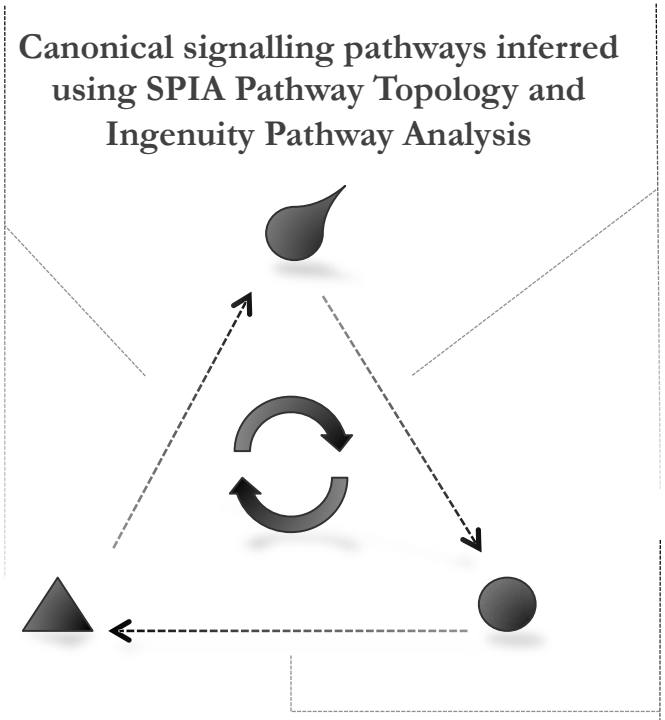


Figure 3.14: Schematic diagram. Differentially expressed genes from tenocyte comparisons used for pathway topology analysis and pathway prediction. Pathways are predicted as activated (+) or inhibited (-). Arrows indicated the direction of the comparison. Most significantly perturbed KEGG pathways are shown; Ingenuity (IPA) canonical pathways enriched with the same data sets are shown. Common upstream regulators are predicted by IPA (box).

|IPA

- Granulocyte adhesion & diapedesis,
- NRF-2 mediated oxidative stress response;
- **Hepatic stellate cell activation**
- Inhibition of matrix metalloproteinases;

|KEGG

+ Cell cycle;
+ p53 signalling pathway;
+ Focal adhesion.

- Transcriptional mis-regulation in cancer;
- **PI3K-Akt signalling pathway**;
- Axon guidance/

|IPA

- **Hepatic stellate cell activation**;
- NRF2-mediated oxidative stress response;
- Role of macrophages, fibroblasts and endothelial cells in rheumatoid arthritis.

|KEGG

+ **PI3K-AKT signalling**;
+ Chemokine signalling pathway;
+ Cytokine-cytokine receptor interaction.

- Cell cycle;
- Transcriptional mis-regulation in cancer;
- Focal adhesion.

Tenocytes

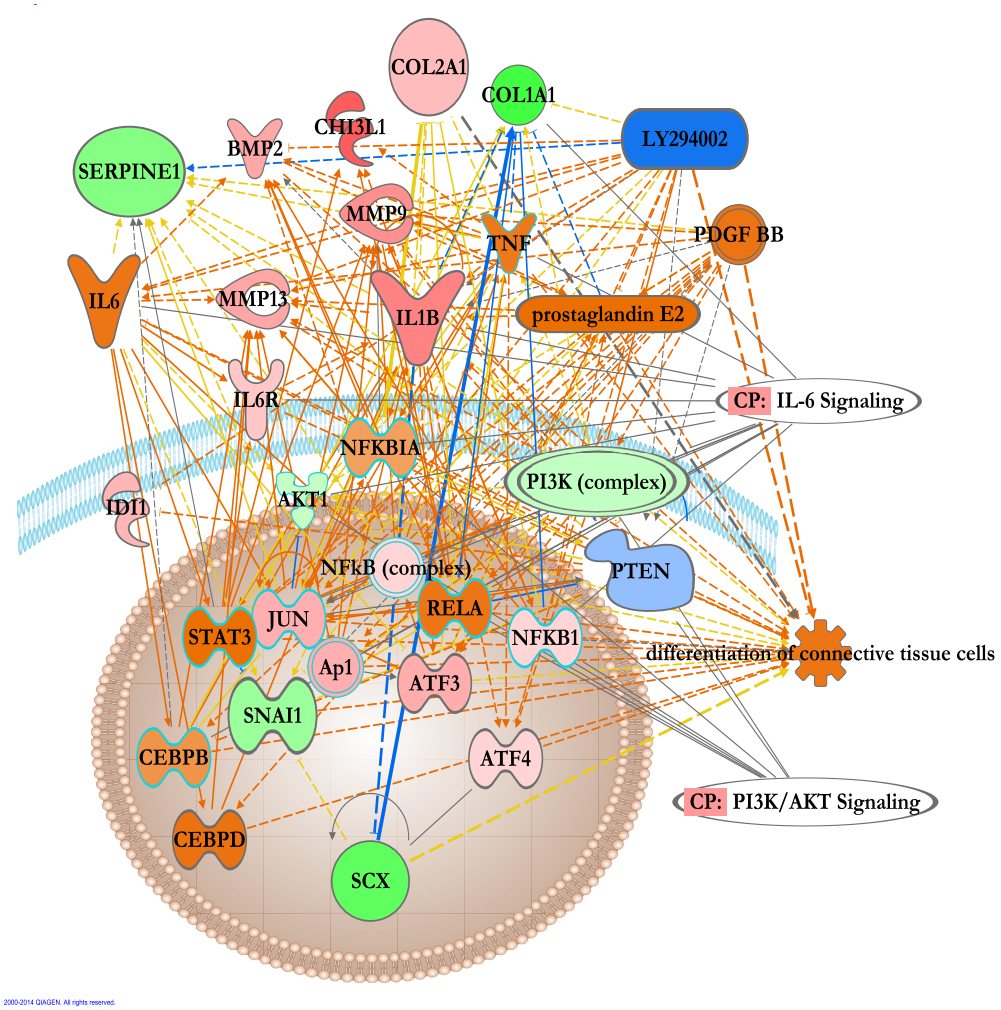
In tenocytes derived from native tendon tissue there was remarkable overlap with chondrocytes in the predicted activated canonical pathways, **Figure 3.14**. The 'PI-3K/Akt pathway' was predicted as activated in the dedifferentiation transition, plus 'cytokine-cytokine receptor interaction'. For the re-differentiation transition there was predicted activation of: 'MAPK signalling pathway', 'transcriptional mis-regulation in cancer', 'ECM receptor signalling pathway', but inhibition of the 'PI-3K/Akt pathway'. In fibrin cultures the expression profile relative to native tissue predicted activation of the canonical pathways 'cell cycle', 'p53 signalling pathway', 'focal adhesion' and 'FOXO signalling pathway'. In contrast to the same comparison in chondrocytes there was predicted inhibition of the 'PI-3K/Akt pathway'. Full lists in **SD3.19**.

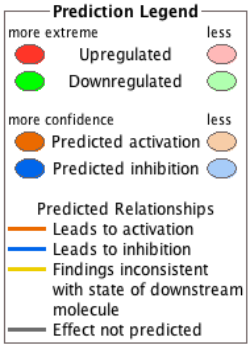
3.3.8: Ingenuity® Pathway Analysis confirms PI-3K activation in mechanistic networks

As previously described, Ingenuity® Pathway Analysis (IPA) was used to further infer key regulators and mechanistic networks for the Affymetrix dataset. Ingenuity® analysis defined several master regulators, the activation status of which was consistent with the differential expression analysis provided from different pairwise comparisons, **Figure 3.15-3.16**. To explore the PI-3K signalling pathway predictions made using pathway topology in **3.3.7** two highly significant upstream regulators were chosen from IPA analysis, the PI-3K activator PDGF BB and the small-molecule inhibitor of PI-3K signalling, LY294002. The PI-3K signalling pathway was predicted to be active in the native to monolayer comparison, whilst inhibited in the monolayer to three-dimensional culture comparison for both chondrocytes and tenocytes.

For the dedifferentiation transition, **Figure 3.15**, the differential gene expression profiles were consistent with PEDF BB activation in chondrocytes (z-score = 3.64, p = 2.4e-22) and tenocytes (z-score = 3.86, p = 4.7e-39). The expression profiles of chondrocytes (z-score = -1.68, p = 1.8e-18) and tenocytes (z-score = -3.14, p = 1.7e-23) from native to monolayer were inconsistent with the suppression of the PI-3K pathway by inhibitor LY294002, i.e. consistent with PI-3K pathway activation in native cells. The activation of the PI-3K pathway in native chondrocytes and tenocytes concur with the findings from pathway topology analysis.

In the re-differentiation transition, **Figure 3.16**, monolayer to three-dimensional cultures, for both cell types, pathway inhibition by LY294002 was predicted, indicated by positive z-scores: chondrocytes (z-score = 3.5, p = 6.8e-17), tenocytes (z-score = 5.19, p = 5.98e-26). The converse was true for PEDF BB with inhibition predicted in chondrocytes (z-score = -4.78, p = 4.56e-30) and tenocytes (z-score = -6.85, p = 1.04e-48). These findings supported PI-3K pathway inhibition in monolayer cells and the converse in cells in three-dimensional cultures.





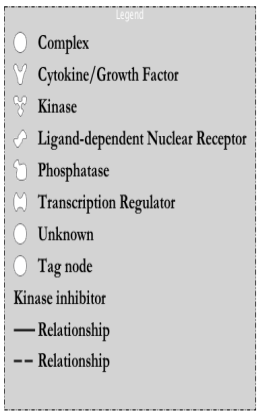
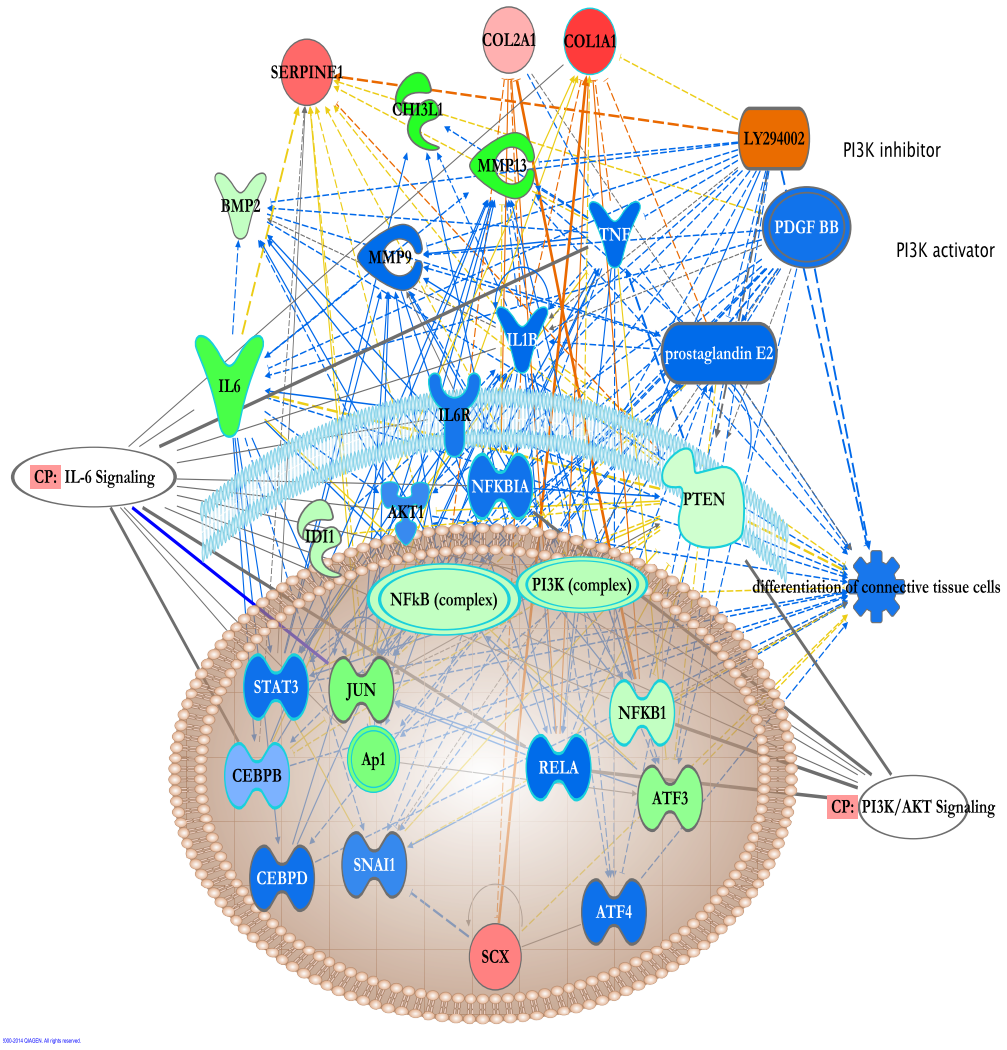
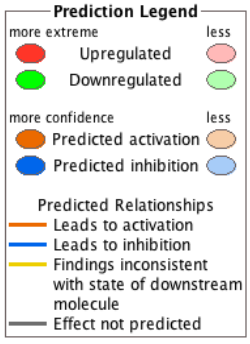


Figure 3.15: IPA Mechanistic Network:
Using the PI-3K small molecule inhibitor LY294002 (see **Figure 3.1**) as the core upstream regulator the mechanistic network presented is coded with differential expression values for the dedifferentiation transition from native chondrocytes to monolayer chondrocytes. Here up-regulated genes (red) represent higher expression in native chondrocytes. The figure legend describes the predicted activation status for other genes in the network. Ingenuity Pathway Analysis predicts that the gene expression profile is consistent with PI-3K signalling activation, i.e. LY294002 not applied. The PI-3K activator PDGF BB is predicted to be active. In this mechanistic network native chondrocytes have higher expression of IL-1B, JUN, ATF3, and *NFκB*. Components of the PI-3K complex show lower express and the PI-3K inhibitor PTEN is predicted to be inhibited. The most significant functional annotation 'differentiation of connective tissue cells' (p=2.4e-26, 24 genes) shows 'activation' where genes native chondrocytes are more highly expressed consistent with the differentiated state.





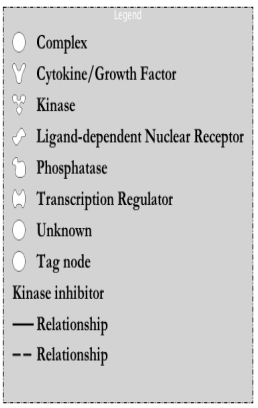


Figure 3.16: IPA Mechanistic Network:
Using the PI-3K small molecule inhibitor LY294002 as the core upstream regulator the mechanistic network presented is coded with differential expression values for the re-differentiation transition from monolayer chondrocytes to alginate beads. Here up-regulated genes (red) represent higher expression in monolayer chondrocytes. The figure legend describes the predicted activation status for other genes in the network. Ingenuity Pathway Analysis predicts that the gene expression profile is consistent with PI-3K signalling inhibition through the application of LY294002. The PI-3K activator PDGF BB is predicted to be inactive. In this mechanistic network chondrocytes in alginate beads have higher expression of IL-6, JUN, ATF3, components of the PI-3K complex and the PI-3K inhibitor PTEN. The functional annotation 'differentiation of connective tissue cells' (p=2.4e-26, 24 genes) shows 'inhibition' where genes in monolayer are more highly expressed.

3.4: Discussion

3.4.1: Study design and rationale

The experimental design in **Chapter 2** compared three three cellular environments – native tissue, passage five monolayer and three-dimensional culture systems – for chondrocytes and tenocytes. Useful comparisons that would elucidate core regulatory pathways in de- and re-differentiation were confounded by the complexity of tissue-specific matricellular components and heterogenous cell populations, e.g. blood cells. To resolve this complexity extra- and peri-cellular matrix was depleted by enzymatic digestion in all conditions. This intervention also acted as a relevant system perturbation as loss of matrix components is common to degenerative disease in cartilage and tendon. Furthermore, underpowered group sizes for some conditions in **Chapter 2** (fibrin cultures, n=2) could not be resolved as production of the Illumina microarray platform ceased. Consequently a novel microarray platform was employed.

Key findings

Gene expression analysis demonstrated considerable overlap in differentially expressed genes between chondrocytes and tenocytes during de- and re-differentiation transitions suggesting a common regulatory pathway could be present. The well-annotated Affymetrix array defined extensive development-associated gene expression in monolayer and three-dimensional cultures for chondrocytes and tenocytes. There were fewer differentially expressed genes found between native cells than between three-dimensional culture-derived cells. By comparing differentially expressed genes common to cartilage or tendon in both Affymetrix and Illumina data sets it was possible to refine a list of tissue-associated genes beyond those found in the published literature. Pathway

topology analysis of matrix-depleted chondrocytes and tenocytes from three conditions defined PI-3K/Akt signalling as the common regulatory pathway, active in native cells and three-dimensional cultures, and inhibited in monolayer culture.

3.4.2: Development-associated genes show higher expression in monolayer

Evidence was presented in **Chapter 2** indicating a trend toward the higher expression of musculoskeletal development-associated genes during dedifferentiation, but this was not consistent and several key genes were not found in the microarray gene lists. In this independent dataset there is further support for the statement that dedifferentiated cells shows a gene expression profile associated with musculoskeletal developmental stages.

At passage three differential expression analysis demonstrated higher expression of scleraxis (*Scx*), Mustang (*Mustn1*) and mohawk (*Mkx*) gene transcripts in both chondrocytes and tenocytes. The author alludes to the common origin of limb tendon and cartilage progenitor cells presented in the **Chapter 1** and the relevance of scleraxis and mohawk to tenogenesis. The role of Mustang is now well-described in chondrogenesis (Lombardo, Komatsu et al. 2004, Gersch and Hadjiargyrou 2009) and has recently been demonstrated to have a role in the regulation of differentiation of myoblasts (Liu, Gersch et al. 2010, Krause, Moradi et al. 2013) likely to be under AP-1 (Fos-Jun dimer) transcriptional activation (Liu and Hadjiargyrou 2006). Given the co-expression of Mustang in areas of chondro- and myo-genesis the expression profile should be further investigated in tendon. In **Chapter 2** significantly higher expression of *Mustn1* was shown in fore- and hind-limb tendons relative to hip cartilage from adult rats. In this

Affymetrix dataset higher expression is also noted in native tendon relative to native cartilage. This is consistent with previously published findings of high expression of Mustang in adult skeletal muscle and tendon (Lombardo, Komatsu et al. 2004). Whilst further studies have begun to elucidate the relevance of Mustang to cartilage and muscle differentiation further work, including confirming marker status in adult tissue, is required for tendon.

3.4.3: Identification of genes expression profiles unique to native tenocytes and chondrocytes

Other than markers of tendon progenitors such as scleraxis, or mature tendon, tenomodulin (*Tnmd*), few tissue-specific markers exist for tendon. Using a consensus differential expression list from the Illumina and Affymetrix data genes with comparable fold changes were considered. Using this approach it was possible to begin to deconstruct the expression profiles defined in **Chapter 2**, which were confounded by heterogenous cell populations. Of the most consistently highly expressed genes, *Tnmd*, *Igfbp6*, *Prrx1*, *Ker*, and *Aspn*, many have been described elsewhere as being strongly representative of tendon relative to other tissues (Jelinsky, Archambault et al. 2010). This analysis contributes several novel candidates for tissue-associated genes, exhibiting consistently higher expression in tendon, for further validation including *Serpinf1*/*Pedf*, *Cxcl13*, *Ecm1* and *Cpxm2*.

Mutations in *Serpinf1*, which encodes pigment epithelium-derived factor (PEDF), are responsible for the phenotype of osteogenesis imperfect (OI) type VI (Homan, Rauch et al. 2011). The spectrum of genetic disorders resulting in the phenotype of bone fragility in OI is derived from mutations affecting the structure and/or synthesis of Type I pro-collagen. However, OI type VI arises from defects in

normal bone mineralization and a recent study suggested that type I collagen formation is normal in these patients. PEDF is collagen binding and has interactions with multiple extracellular matrix components including heparin sulphate proteoglycans (HSPG) and hyaluron. The PEDF-binding motif in the collagen triple helix is overlapping with that of heparin/HSPG and is competed by it; these motifs also localize to the covalent cross-linking sites between the collagen molecules (Sekiya, Okano-Kosugi et al. 2011). It is likely that these genes have key physiological actions in these cells and may not represent tissue-specific markers, e.g. *Serpinf1* shows higher expression in cultured chondrocytes that in native chondrocytes. This study supports the higher expression of *Serpinf1* in isolated native tenocytes (Affymetrix) and whole tendon (Illumina), however, the physiological role of PEDF in mature tendon has yet to be elucidated.

For cartilage, given the greater availability of data defining novel, or candidate, tissue markers is less likely, however, adding to the reference of healthy cartilage expression profiles is relevant. For example, the high expression of *Mmp8* (neutrophil collagenase) is found; together with *Mmp13*, *Mmp8* forms the collagenase superfamily. *Mmp13* is well described as a mediator of cartilage degradation, however, *Mmp8* is expressed in chondrocyte development and in mature articular cartilage (Sasano, Zhu et al. 2002), and may have a protective role in arthritis (García, Forteza et al. 2010).

3.4.4: The PI-3 kinase/Akt pathway in de- and re-differentiation is a common to chondrocytes and tenocytes

In **Chapter 2** there was no clear prevailing signaling pathway defined for de- and re-differentiation transitions using a pathway topology approach. This may have

arisen due to the inherent matricellular differences and complexity of the samples from native tissue.

As described in **Chapter 1** common approaches to defining the signalling pathways resulting in a particular phenotype have relied on two standard statistical techniques, over-representation analysis and functional class scoring. Both of these techniques are limited by the fact that they fail to account for dependencies and interactions at a systems level; by considering each element of a pathway independently these techniques do not provide a unified understanding of the system (Khatri, Draghici et al. 2007). For both the Illumina and Affymetrix data sets an impact analysis method was employed to determine the most perturbed signalling pathways in a given context. This methodology combines the magnitude of gene expression changes, statistical analysis of a set of pathway genes and incorporates knowledge of the pathway topology, signalling interactions and position of differentially expressed genes within a pathway. Against traditional techniques impact factor analysis has been shown to be more sensitive and specific (Tarca, Draghici et al. 2009).

In this study the PI-3K/Akt signalling pathway was shown to be active in native cells and inhibited in monolayer cultures. The activation status was conflicted when alginate or fibrin cultures were compared to their respective native cells using SPIA. However, the pathway was considered to be active in both three-dimensional cultures when a restricted expression profile was analysed using Ingenuity® Pathway Analysis. It is possible that within three-dimensional cultures a re-activation of PI-3K signaling may occur, associated with a redifferentiation phenotype, that may result in conflicted categorical predictions of pathway status

when expression profiles are compared to native cells where the pathway is also active.

The PI-3K/Akt signalling pathway has been implicated in a range of physiological activities including dedifferentiation, proliferation, matrix synthesis and cell survival (Beier and Loeser 2010). Cellular effects are regulated through a broad range of downstream targets including mTOR, NF-κB, GSK-3B and p53 (Chen, Crawford et al. 2013).

The promotion of matrix synthesis and survival by PI-3K signalling in chondrocytes has been demonstrated in a number of studies. Enhanced proteoglycan synthesis by IGF1-mediated stimulation of chondrocytes, either in monolayer or alginate beads, is PI-3K/Akt/mTOR-mediated, possibly by promotion of translational activity (Starkman, Cravero et al. 2005). Constitutive expression of Akt in human articular chondrocytes resulted in the significant increase of proteoglycan synthesis and elevated *Sox9* and *Col2a1* expression (Yin, Park et al. 2009); in contrast, oxidative stress is shown to inhibit these IGF1-induced effects (Oh and Chun 2003, Yin, Park et al. 2009). Recently, the negative regulator of PI-3K signaling *Pten*, when down-regulated by siRNA, was shown to result in an increase in the expression of hallmarks of differentiated cartilage and promoted proteoglycan synthesis under oxidative stress conditions (Iwasa, Hayashi et al. 2014), thereby confirming the pro-matrix effects of PI-3K signalling.

Much of our understanding of PI-3K effects on chondrocyte differentiation arises from the study of endochondral ossification. PI-3K signalling is critical for long bone formation, as demonstrated by *Akt* knockout mice (Ulici, Hoenselaar et al. 2009) and in the terminal differentiation of chondrocytes in the growth plate

(Ulici, Hoenselaar et al. 2008). A number of studies implicate the wider PI-3K pathway as being of relevance in the differentiation status of chondrocytes or MSCs (Chen, Crawford et al. 2013).

The manner by which PI-3K signalling is involved in the balance of chondrocyte survival and terminal differentiation is not clear. Kita, *et al* (2008), reported that conditional *Akt* activation in organ cultures suppressed the expression of markers of chondrocyte hypertrophy whilst increasing proliferation. Furthermore, inhibition of PI-3K signalling increased chondrocyte terminal differentiation and resulted in reduced bone length in a an embryonic fore-limb culture (Kita, Kimura et al. 2008). Reduced bone length, proliferative and hypertrophic growth plate zones were also reported by Ulici, *et al* (2008) in response to pharmacological inhibition of PI-3K by LY294002 (Ulici, Hoenselaar et al. 2008); this study reported that PI-3K inhibition suppressed early and late markers of chondrocyte differentiation, rather than enhancing hypertrophic differentiation. Phosphorylated Akt was found, however, throughout the late proliferative and early hypertrophic chondrocytes in the growth plate. Ikegami, *et al* (2011) found that constitutive expression of *Sox9* in the developing mouse growth plate was essential for chondrocyte survival and subsequent hypertrophy by promotion of Akt phosphorylation (Ikegami, Akiyama et al. 2011) supporting a pro-hypertrophy role for PI-3K signalling. This study also stated that PI-3K/Akt signalling retarded the transition to hypertrophic differentiation from the proliferative state during endochondral ossification; temporal expression of *Runx2* would, however, drive terminal differentiation in concert with PI-3K signalling.

Comparison of *in silico* predictions in this study with previous *in vitro* work is problematic. The studies above consider the findings of their investigations in

isolated terms and so also the status of the pathway under investigation is defined by binary expression of individual elements. Pathway topology techniques define, on the basis of global perturbations, the activation status across multiple pathways.

Novel findings in this study, employing pathway topology analysis using canonical KEGG pathways and inference of upstream regulators by Ingenuity® Pathway Analysis, supported activation of PI-3K signalling in native chondrocytes and tenocytes. This is consistent with a pro-differentiation, pro-matrix role for PI-3K signalling. In contrast, monolayer culture was associated with a predicted inhibition of PI-3K activity. In both native chondrocytes and tenocytes this analysis shows the increased expression of PI-3K inhibitor *Pten* phosphatase relative to monolayer culture, but also higher expression of *Pten* in monolayer chondrocytes relative to alginate cultures. If the activation status of the PI-3K pathway were considered *prima facie* this would imply that the pathway was inhibited in both transitions by *Pten* expression. However, because of the inclusion thresholds set by IPA the expression levels of *Pten* in native chondrocytes and tenocytes were considered too low and this exclusion may influence the activation status predicted. Clearly validation of the mRNA and protein levels of PI-3K pathway components is required.

The interplay between PI-3K and other pathways requires further discussion. In the mechanistic networks generated by IPA the top upstream regulators were TGF-β, IL-6 and IL-1β. It is possible to rationalize these predicted regulators with known PI-3K signalling interactions.

In human articular chondrocytes phosphorylation of Akt has been found following exposure to IL-6 (plus soluble IL-6R) or oncostatin M (OSM)

(Litherland, Dixon et al. 2008). In contrast IL-1 exposure alone did not result in Akt activation. The PI-3K inhibitor LY294002 inhibited cartilage degradation in the presence of IL-1+OSM with a concentration dependent reduction in the induction of MMP1 and MMP13. Given the strong evidence for the role of mediators such as IL-6 and OSM in cartilage ECM degradation the role of PI-3K/Akt in cartilage destruction is relevant. Shakibaei, *et al* (2007) presented evidence to support inhibition of phosphorylation of Akt by curcumin as a mechanism to suppress Il-1β -mediated up-regulation of COX2 and MMP9 (Shakibaei, John et al. 2007).

The balance of anabolic or catabolic processes in cartilage in response to PI-3K signaling may be ligand specific (Beier and Loeser 2010). TGF-β, by binding to cell surface receptor serine/threonine kinases, initiates signal transduction via SMAD and non-SMAD signaling pathways; transduction through the PI-3K/Akt/mTOR pathway has been described for chondrocytes. Qureshi, *et al* (2007) found that TGF-β stimulation of human chondrocytes activated Akt in a delayed manner; this induction was suppressed by PI-3K and Akt inhibitors (Qureshi, Ahmad et al. 2007). Rapamycin, an inhibitor of mTOR, suppressed transcription of the tissue inhibitor of metalloproteinases-3, TIMP3, in response to TGF-β stimulation. This suggested that, in part, TGF-β signaling occurred through pro-matrix PI-3K/Akt/mTOR signalling. Recently signaling through Akt/mTOR, following TGF-β activation, was required for the expression of chondrogenesis-associated genes in mouse pre-cartilaginous stem cells (Li, Wang et al. 2014).

Whilst it must be recognised that the prevailing signals resulting in these gene expression profiles arise from the interplay and convergence of multiple signalling

pathways the dominance of PI-3K/Akt signalling in this study provides a strong rational for further investigation.

In conclusion, the PI-3K/Akt signaling pathway has established effects on the differentiation, matrix synthesis, survival, and homeostasis of chondrocytes; additionally key roles in the regulation of inflammatory mediators and downstream effector pathways, such as NF-κB, implicates PI-3K/Akt signaling further in chondrocyte dysregulation. Some intervention studies have suggested modulation of this pathway as a mechanism for the treatment of osteoarthritis (Chen, Crawford et al. 2013). In a study of tendon-derived stem cells PI-3K signaling was associated with a promotion of osteogenic differentiation implicating this signaling cascade in tendon calcification (Liu, Chen et al. 2013). In this chapter evidence suggests involvement of PI-3K/Akt signalling in the mediation of de- and re-differentiation for both chondrocytes and tenocytes. Clearly the complexity of the pathway is lost in categorical descriptions of pathway activation – the ratio of PI-3K isoforms, the Akt subunit targeted, the downstream effector signaling cascade, and activation state/phosphorylation, cannot be interpreted from this data. Further work with immunoblotting, qPCR, small-molecule inhibition and further genome-wide transcriptome analysis would be required to validate these findings.

3.4.4: Data and study limitations

In line with the majority of microarray studies investigations the study presented here is likely to be underpowered, with four biological replicates in each group. The use of isogenic rats and paired tissue samples should mitigate this to some extent by reducing the signal-to-noise ratio. Although considerable overlap is demonstrated in differential gene expression between the Illumina and Affymetrix studies the greatest correlation is evident only for highly differentially expressed

genes. That comparable expression profiles are evident across two microarray platforms increases the confidence of these findings. This comparability also goes some way to explain that the close association in expression profiles between matrix-depleted native chondrocytes and tenocytes in this study is not a response to extended enzymatic digestion protocols alone.

It is not unexpected that a strong comparison in pathway prediction between the Illumina and Affymetrix datasets is not found. In the former monolayer cells were considered at passage five, the latter at passage three. Chondrocytes in alginate beads were only retained *in situ* for seven days to match the total culture period for fibrin constructs; in the Illumina study alginate beads spent two weeks in culture. Furthermore, by depleting the matrix in this study the sample complexity was reduced relative to the Illumina study.

Methodological and conceptual evolution of systems biology is still required. Issues associated with pathway analysis include the lack of consistent appraisal, no consensus on the representation of pathways in knowledge-bases, varied definitions of a pathway, failure to reproduce pathway analysis results, and the lack of integration with biochemical models, lack of condition or tissue-specificity, or inclusion of time-dependent models, for example, developmental stages (Mitrea, Taghavi et al. 2013, Conesa and Mortazavi 2014). Most functional annotation databases are not dynamic, rather they are biased by the weight of publications toward high impact research topics such as cancer and neurodegenerative disease. The boundaries of pathway definitions and gene annotations are often arbitrary, incomplete, and the functional assignment of genes is often redundant or nebulous. However, methodology presented here goes some way to deconstructing the gene expression profile associated with *in vivo* complexity and

isolating the regulatory pathways involved in chondrocyte and tenocyte responses to changes in the three-dimensional environment.

References

Beier, F. and R. Loeser (2010). "Biology and pathology of Rho GTPase, PI-3 kinase-Akt, and MAP kinase signaling pathways in chondrocytes." Journal of Cellular Biochemistry **110**(3): 573-580.

Cantley, L. (2002). "The phosphoinositide 3-kinase pathway." Science (New York, N.Y.) **296**(5573): 1655-1657.

Carvalho, B. and R. Irizarry (2010). "A framework for oligonucleotide microarray preprocessing." Bioinformatics (Oxford, England) **26**(19): 2363-2367.

Chen, J., R. Crawford, C. Chen and Y. Xiao (2013). "The key regulatory roles of the PI3K/Akt signaling pathway in the functionalities of mesenchymal stem cells and applications in tissue regeneration." Tissue Engineering. Part B, Reviews **19**(6): 516-528.

Chen, J., R. Crawford and Y. Xiao (2013). "Vertical inhibition of the PI3K/Akt/mTOR pathway for the treatment of osteoarthritis." Journal of Cellular Biochemistry **114**(2): 245-249.

Conesa, A. and A. Mortazavi (2014). "The common ground of genomics and systems biology." BMC Systems Biology **8 Suppl 2**.

Cox, B. and A. Emili (2006). "Tissue subcellular fractionation and protein extraction for use in mass-spectrometry-based proteomics." Nature Protocols **1**(4): 1872-1878.

Fonslow, B., B. Stein, K. Webb, T. Xu, J. Choi, S. Park and J. Yates (2013). "Digestion and depletion of abundant proteins improves proteomic coverage." Nature Methods **10**(1): 54-56.

Foster, J., M. Blunt, E. Carter and S. Ward (2012). "Inhibition of PI3K Signaling Spurs New Therapeutic Opportunities in Inflammatory/Autoimmune Diseases and Hematological Malignancies." Pharmacological Reviews **64**(4): 1027-1054.

García, S., J. Forteza, C. López-Otin, J. Gómez-Reino, A. González and C. Conde (2010). "Matrix metalloproteinase-8 deficiency increases joint inflammation and bone erosion in the K/BxN serum-transfer arthritis model." Arthritis Research & Therapy **12**(6): R224.

Gersch, R. and M. Hadjiargyrou (2009). "Mustn1 is expressed during chondrogenesis and is necessary for chondrocyte proliferation and differentiation in vitro." Bone **45**(2): 330-338.

Homan, E., F. Rauch, I. Grafe, C. Lietman, J. Doll, B. Dawson, T. Bertin, D. Napierala, R. Morello, R. Gibbs, L. White, R. Miki, D. Cohn, S. Crawford, R. Travers, F. Glorieux and B. Lee (2011). "Mutations in SERPINF1 cause osteogenesis imperfecta type VI." Journal of Bone and Mineral Research **26**(12): 2798-2803.

Ikegami, D., H. Akiyama, A. Suzuki, T. Nakamura, T. Nakano, H. Yoshikawa and N. Tsumaki (2011). "Sox9 sustains chondrocyte survival and hypertrophy in part through Pik3ca-Akt pathways." Development (Cambridge, England) **138**(8): 1507-1519.

Irizarry, R., B. Hobbs, F. Collin, Y. Beazer - Barclay, K. Antonellis, U. Scherf and T. Speed (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics **4**(2): 249-264.

Iwasa, K., S. Hayashi, T. Fujishiro, N. Kanzaki, S. Hashimoto, S. Sakata, N. Chinzei, T. Nishiyama, R. Kuroda and M. Kurosaka (2014). "PTEN regulates matrix synthesis in adult human chondrocytes under oxidative stress." Journal of Orthopaedic Research **32**(2): 231-237.

Jelinsky, S., J. Archambault, L. Li and H. Seeherman (2010). "Tendon-selective genes identified from rat and human musculoskeletal tissues." Journal of Orthopaedic Research **28**(3): 289-297.

Khatri, P., S. Draghici, A. L. Tarca, S. S. Hassan and R. Romero (2007). A System Biology Approach for the Steady-State Analysis of Gene Signaling Networks. Progress in Pattern Recognition, Image Analysis and Applications. **4756:** 32-41.

Kita, K., T. Kimura, N. Nakamura, H. Yoshikawa and T. Nakano (2008). "PI3K/Akt signaling as a key regulatory pathway for chondrocyte terminal differentiation." Genes to Cells **13**(8): 839-850.

Krause, M. P., J. Moradi, S. K. Coleman, D. M. D'Souza, C. Liu, M. S. Kronenberg, D. W. Rowe, T. J. Hawke and M. Hadjiargyrou (2013). "A novel GFP reporter mouse reveals Mustn1 expression in adult regenerating skeletal muscle, activated satellite cells and differentiating myoblasts." Acta Physiologica (Oxford, England) **208**(2): 180-190.

Li, C., Q. Wang and J.-F. Wang (2014). "Transforming growth factor-β (TGF-β) induces the expression of chondrogenesis-related genes through TGF-β receptor II (TGFRII)-AKT-mTOR signaling in primary cultured mouse precartilaginous stem cells." Biochemical and Biophysical Research Communications **450**(1): 646-651.

Litherland, G., C. Dixon, R. Lakey, T. Robson, D. Jones, D. Young, T. Cawston and A. Rowan (2008). "Synergistic Collagenase Expression and Cartilage Collagenolysis Are Phosphatidylinositol 3-Kinase/Akt Signaling-dependent." Journal of Biological Chemistry **283**(21): 14221-14229.

Liu, C., R. Gersch, T. Hawke and M. Hadjiargyrou (2010). "Silencing of Mustn1 inhibits myogenic fusion and differentiation." American Journal of Physiology: Cell Physiology **298**(5): C1100-1108.

Liu, C. and M. Hadjiargyrou (2006). "Identification and characterization of the Mustang promoter: regulation by AP-1 during myogenic differentiation." Bone **39**(4): 815-824.

Liu, J., L. Chen, X. Tao and K. Tang (2013). "Phosphoinositide 3-kinase/Akt signaling is essential for prostaglandin E2-induced osteogenic differentiation of rat tendon stem cells." Biochemical and Biophysical Research Communications **435**(4): 514-519.

Lombardo, F., D. Komatsu and M. Hadjiargyrou (2004). "Molecular cloning and characterization of Mustang, a novel nuclear protein expressed during skeletal development and regeneration." The FASEB Journal **18**(1): 52-61.

Mastrokolias, A., J. den Dunnen, G. van Ommen, P. Hoen and M. Willeke van Roon (2012). "Increased sensitivity of next generation sequencing-based expression profiling after globin reduction in human blood RNA." BMC Genomics **13**(1): 28.

McNamara, C. and A. Degterev (2011). "Small-molecule inhibitors of the PI3K signaling network." Future Medicinal Chemistry **3**(5): 549-565.

Mitrea, C., Z. Taghavi, B. Bokanizad, S. Hanoudi, R. Tagett, M. Donato, C. Voichiţa and S. Drăghici (2013). "Methods and approaches in the topology-based analysis of biological pathways." Frontiers in Physiology **4**: 278.

Nygaard, V. and E. Hovig (2008). Cell sampling and global nucleic acid amplification. Practical Systems Biology. A. Hetherington and C. Grierson, Taylor and Francis Group. **61:** 17-34.

Oh, C.-D. and J.-S. Chun (2003). "Signaling mechanisms leading to the regulation of differentiation and apoptosis of articular chondrocytes by insulin-like growth factor-1." The Journal of Biological Chemistry **278**(38): 36563-36571.

Pearson, R., X. Liu, G. Sanguinetti, M. Milo, N. Lawrence and M. Rattray (2009). "puma: a Bioconductor package for propagating uncertainty in microarray analysis." BMC Bioinformatics **10**(1): 211.

Qureshi, H. Y., R. Ahmad, J. Sylvester and M. Zafarullah (2007). "Requirement of phosphatidylinositol 3-kinase/Akt signaling pathway for regulation of tissue inhibitor of metalloproteinases-3 gene expression by TGF-beta in human chondrocytes." Cellular Signalling **19**(8): 1643-1651.

Ritty, T., R. Roth and J. Heuser (2003). "Tendon cell array isolation reveals a previously unknown fibrillin-2-containing macromolecular assembly." Structure (London, England : 1993) **11**(9): 1179-1188.

Sasano, Y., J.-X. Zhu, M. Tsubota, I. Takahashi, K. Onodera, I. Mizoguchi and M. Kagayama (2002). "Gene expression of MMP8 and MMP13 during embryonic development of bone and cartilage in the rat mandible and hind limb." The Journal of Histochemistry and Cytochemistry **50**(3): 325-332.

Sekiya, A., H. Okano-Kosugi, C. Yamazaki and T. Koide (2011). "Pigment epithelium-derived factor (PEDF) shares binding sites in collagen with heparin/heparan sulfate proteoglycans." The Journal of Biological Chemistry **286**(30): 26364-26374.

Shakibaei, M., T. John, G. Schulze-Tanzil, I. Lehmann and A. Mobasheri (2007). "Suppression of NF-kappaB activation by curcumin leads to inhibition of expression of cyclo-oxygenase-2 and matrix metalloproteinase-9 in human articular chondrocytes: Implications for the treatment of osteoarthritis." Biochemical Pharmacology **73**(9): 1434-1445.

Sims, D., I. Sudbery, N. Ilott, A. Heger and C. Ponting (2014). "Sequencing depth and coverage: key considerations in genomic analyses." Nature Reviews: Genetics **15**(2): 121-132.

Starkman, B., J. Cravero, M. Delcarlo and R. Loeser (2005). "IGF-I stimulation of proteoglycan synthesis by chondrocytes requires activation of the PI 3-kinase pathway but not ERK MAPK." The Biochemical Journal **389**(Pt 3): 723-729.

Tarca, A. L., S. Draghici, P. Khatri, S. Hassan, P. Mittal, J.-S. Kim, C. J. Kim, J. P. Kusanovic and R. Romero (2009). "A novel signaling pathway impact analysis." Bioinformatics (Oxford, England) **25**(1): 75-82.

Ulici, V., K. Hoenselaar, H. Agoston, D. McErlain, J. Umoh, S. Chakrabarti, D. Holdsworth and F. Beier (2009). "The role of Akt1 in terminal stages of endochondral bone formation: angiogenesis and ossification." Bone **45**(6): 1133-1145.

Ulici, V., K. Hoenselaar, R. Gillespie and F. Beier (2008). "The PI3K pathway regulates endochondral bone growth through control of hypertrophic chondrocyte differentiation." BMC Developmental Biology **8**(1): 40.

Vanhaesebroeck, B., L. Stephens and P. Hawkins (2012). "PI3K signalling: the path to discovery and understanding." Nature Reviews. Molecular Cell Biology **13**(3): 195-203.

Wilson, R., J. Whitelock and J. Bateman (2009). "Proteomics makes progress in cartilage and arthritis research." Matrix Biology **28**(3): 121-128.

Yin, W., J.-I. Park and R. Loeser (2009). "Oxidative stress inhibits insulin-like growth factor-I induction of chondrocyte proteoglycan synthesis through differential regulation of phosphatidylinositol 3-Kinase-Akt and MEK-ERK MAPK signaling pathways." The Journal of Biological Chemistry **284**(46): 31972-31981.

Appendix 3

R Codes

Processing and analysis of raw Affymetrix microarray data

```
setwd("/Users/XXX")

library(oligo)
library(pd.ragene.2.0.st)
library(ragene20sttranscriptcluster.db)
#library(ragene20stprobeset.db)

     cels<-list.celfiles(full.names=TRUE)
     affymetrix.data<-read.celfiles(cels)

     affy.corrected<-
     rma(affymetrix.data,normalize=FALSE,target="core")

##########################################################
####NON-SPECIFIC FILTERING#################################
##########################################################

library(genefilter)

     filter<-
     nsFilter(affy.corrected,require.entrez=FALSE,var.func=IQR,v
     ar.cutoff=0.5,var.filter=TRUE,filterByQuantile=TRUE,
     feature.exclude="^AFFX",remove.dupEntrez=FALSE)

     affy.filtered<-filter$eset
     dim(exprs(affy.filtered))

##########################################################
####LOESS NORMALISATION#################################
##########################################################
##tidy-up and release memory
#rm(cels, affy.corrected,affymetrix.data,filter,affy.filtered)
library(WGCNA)
collectGarbage()
```

```
detach("package:WGCNA")



library(limma)
     affy.loess<-normalizeCyclicLoess(exprs(affy.filtered),
     weights = NULL, span=0.7, iterations = 3, method = "affy")

     save(affy.loess,file="AFFY2014_loess_filtered.RData")
##########################################################
####CONTRAST MATRIX FOR MODEL#############################
##########################################################
     design<-model.matrix(~0+factor(
     c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4,5,5,5,5,6,6,6,6)))
     colnames(design)<-c("ALG","dC","dT","FIB","nC","nT")
     rownames(design)<-c("ALG_1.CEL", "ALG_2.CEL", "ALG_3.CEL" ,
     "ALG_4.CEL" ,"dC_1.CEL" , "dC_2.CEL"  ,"dC_3.CEL",
     "dC_4.CEL" , "dT_1.CEL" , "dT_2.CEL" , "dT_3.CEL",
     "dT_4.CEL" , "FIB_1.CEL" ,"FIB_2.CEL" ,"FIB_3.CEL",
     "FIB_4.CEL", "nC_1.CEL" , "nC_2.CEL"  ,"nC_3.CEL",
     "nC_4.CEL" , "nT_1.CEL" , "nT_2.CEL" ,
     "nT_3.CEL" , "nT_4.CEL")

##########################################################
####ANNOTATION FILE#######################################
##########################################################
     ID<-rownames(affy.loess)
     symbol<-
     mget(ID,ragene20sttranscriptclusterSYMBOL,ifnotfound=NA)
     genename=mget(ID,ragene20sttranscriptclusterGENENAME,
     ifnotfound=NA)
     entrezID=mget(ID,ragene20sttranscriptclusterENTREZID,
     ifnotfound=NA)

     anno=data.frame(Illumina_ID=ID,
     Symbol=as.character(symbol),
     EntrezID=as.numeric(entrezID),
     GeneName=as.character(genename))

##########################################################
####DIFFERENTIAL EXPRESSION###############################
##########################################################
library(limma)

     fit<-lmFit(affy.loess,design)
     contrast.matrix<-makeContrasts(nC-dC,levels=design)
#set up the matrix and then you can include or exclude the
samples that you want
     fit2<-contrasts.fit(fit,contrast.matrix)
     fit2<-eBayes(fit2)
     fit2$gene=anno

     DEgenes=topTable(fit2,coef=1,number=5000,lfc=0.5,
     adjust.method="fdr",sort.by="logFC",genelist=fit2$gene)

     DEgenes$Probe.ID<-rownames(DEgenes)

     write.csv(DEgenes,
     file="Differential_expression_Affy_June2014.csv")
```

```
#########################################################
####COEXPRESSION AND PLOTS###############################
#########################################################
     data<-affy.loess
     write.csv(data,file="data.csv")
     data<-read.csv(file="data.csv",sep=",",header=TRUE)
     colnames(data)[1] <- c("ProbeID")

     ArrayName=names(data.frame(data[,-1]))
     GeneName=data$EntrezID
     exprs=data.frame(t(data[,-1]))
     names(exprs)=data[,1]
     dimnames(exprs)[[1]]=names(data.frame(data[,-1]))
     exprs.v=as.vector(apply(as.matrix(exprs),2,var,na.rm=T))
#calculate variance across the expression data - default Chapter
2 is v>0.2

     present=as.vector(apply(!is.na(as.matrix(exprs)),2,sum))
     keep=exprs.v>0.3 & present>=4
     table(keep)

     filt=exprs[,keep]

library(WGCNA)

     powers = c(c(1:10), seq(from = 12, to=20, by=2))
     # Call the network topology analysis function
     sft = pickSoftThreshold(filt,
     powerVector = powers, verbose = 5)
# Plot the results:
#setwd("/Users/alanmueller/Desktop/Thesis/Chapter_2/Chapter2_ima
ges")
#pdf("Soft_Threshold_Choice.pdf",height=8,width=12)
#sizeGrWindow(9, 5)
#par(mfrow = c(1,2))
cex1 = 0.9


#Scale-free topology fit index as a function of the soft-
thresholding power
     plot(sft$fitIndices[,1], -
     sign(sft$fitIndices[,3])*sft$fitIndices[,2],xlab="Soft
     Threshold (power)",ylab="Scale Free Topology Model
     Fit,signed R^2",type="n",main = paste("Scale
     independence"))
     text(sft$fitIndices[,1], -
     sign(sft$fitIndices[,3])*sft$fitIndices[,2],labels=powers,c
     ex=cex1,col="steelblue2")
# this line corresponds to using an R^2 cut-off of h
     abline(h=0.90,col="grey66")


#####create adjacency matrix
     ADJ1=abs(cor(filt,use="p"))^4
     k=as.vector(apply(ADJ1,2,sum, na.rm=T))
##define connectivities
      sizeGrWindow(10,5)
      par(mfrow=c(1,2))
      hist(k)
      scaleFreePlot(k, main="Check scale free topology\n")
      datExpr=filt[, rank(-k,ties.method="first" )<=500]
```

```
     setwd("/Users/XXX")
     save(datExpr,file="matrix_for_affy2014_PCA_heatmap.RData")

##########################################################
####PCA PLOT##############################################
##########################################################
library(FactoMineR)
library(RColorBrewer)

     scaled<-scale(t(datExpr))
     colnames(scaled)

     res.pca<-PCA(t(scaled),graph=FALSE,axes=c(1,2))##or 'map'

     PC1 <- res.pca$ind$coord[,1]
     PC2 <- res.pca$ind$coord[,2]


     condition<-
     c(rep("3D.Model",4),rep("Monolayer",8),rep("3D.Model",4),re
     p("Native",8))
     condition<-as.data.frame(condition)
     cell.type<-
     c(rep("chondrocyte",8),rep("tenocyte",8),rep("chondrocyte",
     4),rep("tenocyte",4))
     cell.type<-as.data.frame(cell.type)

     PCs <- data.frame(cbind(PC1,PC2,cell.type,condition))

     PCA.comp1<-res.pca$eig[1,2]
     PCA.comp2<-res.pca$eig[2,2]

     #mypalette<-brewer.pal(3,"Greys")
     #or
     mypalette<-c("gray0","gray88","gray64")

library(ggplot2)

     setwd("/Users/XXX")
     pdf(file = "Affymetrix_PCA.pdf", width= 8,
     height = 8,useDingbats=F)
     par(mar=c(1,1,1,1))
     p<-ggplot(PCs)
     p<-
     p+geom_point(aes(PC1,PC2,color=condition,shape=cell.type),
     size=6,alpha=0.6)+
     scale_colour_manual(values=mypalette)+
     labs(list(x=sprintf("PC1(%.1f%%)",PCA.comp1),
     y=sprintf("PC2(%.1f%%)",PCA.comp2)))+
     theme_minimal(base_size=10,base_family="Helvetica")+
     theme(legend.position = "bottom",text =
     element_text(size=12),
     plot.title=element_text(lineheight=.8,face="bold"))+
     ggtitle("Principal Component Analysis\nn=24 Affymetrix Rat
     Gene ST 2.0")+
     scale_shape_discrete(solid=T)
     p
     dev.off()
```

```
#########################################################
####HEATMAP PLOT#########################################
#########################################################
library(gplots)

     matrix<-as.matrix(t(datExpr))ß
     h<-heatmap.2(matrix)
     labels(h$rowDendrogram[[1]])
     labels(h$rowDendrogram[[2]][[2]])

##Prepare row groupings
     setwd("/Users/XXX")
     load(file="matrix_for_affy2014_PCA_heatmap.RData")

     hc.rows<- hclust(dist(matrix))
     plot(hc.rows)
     ct<-cutree(hc.rows,h=15)
     plot(hc.rows)
     rect.hclust(hc.rows,h=15)
     ##have prepared as four groups - this looks the most
     sensible split
     ##number of genes in each group
     table(ct)

     hm <- heatmap.2(matrix)
     hc <- as.hclust(hm$rowDendrogram )
     groups<-cutree(hc, h=25.5) [hc$order]
     names<-names(groups)
     groups1<-unname(groups)
     groups2<-data.frame("Symbol"=names,"Groups"=groups1)

     ##load annotation of the top 500 genes
     data<-
     read.csv(file="500_covariant_genes_PCA_heatmap.csv",header=
     TRUE)
     merged<-merge(data,groups2,by.x="Probe_ID",by.y="Symbol")
     write.csv(merged,file="heatmapGroups_sept2014.csv",row.name
     s=FALSE)

library(gplots)

     matrix<-as.matrix(t(datExpr))

     colnames(matrix)<-
     c("ALGINATE.1","ALGINATE.2","ALGINATE.3","ALGINATE.4","MONO
     LAYER_C.1","MONOLAYER_C.2","MONOLAYER_C.3","MONOLAYER_C.4",
     "MONOLAYER_T.1","MONOLAYER_T.2","MONOLAYER_T.3","MONOLAYER_
     T.4","FIBRIN.1","FIBRIN.2","FIBRIN.3","FIBRIN.4","NATIVE_C.
     1","NATIVE_C.2","NATIVE_C.3","NATIVE_C.4","NATIVE_T.1","NAT
     IVE_T.2","NATIVE_T.3","NATIVE_T.4")

     ##Create heatmap with the row groups coloured too
     library(RColorBrewer)
     groups<-cutree(hc,h=25.5)
     cols <- brewer.pal(max(groups), "Set3")
     #pdf(file = "Affymetrix_PCA.pdf", width= 8, height =
     8,useDingbats=F)
     #library(WGCNA)
     pdf("affy2014_heatmap_Sept.pdf",height=12,width=12)
     #sizeGrWindow(9,9)
     #par(mar=c(10,2,10,2))
```

```
      heatmap.2(matrix,col=greenred(100),symkey=TRUE,trace='none'
      ,density.info='none',RowSideColors=cols[groups],ColSideColo
      rs=c(rep("lightsteelblue1",4),rep("goldenrod3",8),rep("ligh
      tsteelblue3",4),rep("midnightblue",4),rep("firebrick1",4))
      ,cexRow=0.09,cexCol=0.8,scale='row',mar=c(4,4),dendrogram='
      both',colsep=c(4,8,12,16,20),sepcolor="white",sepwidth=c(0.
      05,0.05))
      dev.off()


#####################################################################
##Finding overlaps between DE genes in Illumina and Affymetrix##
#####################################################################
##Need lists of Entrez IDs from each data set
      setwd("/Users/XXX")
      illumina<-read.csv(file="illumina.csv",header=TRUE,sep=",")
      affymetrix<-
      read.csv(file="affymetrix.csv",header=TRUE,sep=",")

      consensus<-merge(
      affymetrix,illumina,by.x="EntrezID",by.y="EntrezID",all.x=T
      RUE)

      ##ensure those unique to affymetrix data are clear
      setwd("/Users/XXX")
      write.csv(consensus,file="Illumina_Affymetrix_consensus.csv
      ")


      ####Plot common log fold changes and correlations
      setwd("/Users/XXX")
      cor<-read.csv(file="test_data.csv",header=TRUE)
      x<-cor$logFC.Affy
      y<-cor$log2FC.Illumina
      cor.test(x,y,method="pearson")

library(ggplots2)

      #mypalette<- see above
      condition<-
      c(rep("Upregulated",168),rep("Downregulated",226))
      setwd("/Users/XXX")
      pdf("nT_dT_Illumina_Affy_correlation.pdf",height=8,width=8)
      g<-ggplot(cor, aes(x=logFC.Affy, y=log2FC.Illumina,
      color=condition, shape=condition),alpha=0.5,size=10) +
      geom_point()+geom_smooth(method=lm,se=FALSE)
      g<-
      g+scale_color_manual(values=mypalette)+scale_shape_discrete
      (solid=T)
      g<-
      g+theme_minimal(base_size=10,base_family="Helvetica")+theme
      (plot.title=element_text(lineheight=.9,face="bold"))
      g<-g+ggtitle("Differentially Expressed Genes Common to
      Illumina and Affymetrix Studies\n Native Chondrocytes vs.
      Dedifferentiated Chondrocytes")
      g<-g+labs(list(x=sprintf("log2 fold change
      Affymetrix"),y=sprintf("log2 fold change Illumina")))
      g<-g+annotate("text",label="r=0.81, p<2.2e-16",x=3,y=-
      2.5,size=4,colour="black")
      g
      dev.off()
```

```
#####################################################################
######Venn Euler Diagrams#############################################
#####################################################################
library(venneuler)

     v<-venneuler(c(A=3863,B=2709,"A&B"=782))

     ##covert the colours to R readable colour-strings
     col.fn <- function(col, alpha=0.3) {
         col<- hcl(col * 360, 130, 60)
         col <- col2rgb(col)/255
         col <- rgb(col[1, ], col[2, ], col[3, ], alpha)
         col
     }
     COL <- col.fn(v$colors)
     LABS <- v$labels
     id <- match(names(v$colors), LABS)

     leg.txt<-c("Affymetrix","Illumina")

     par(font.main=1,font.lab=1,family="Helvetica",mai=c(1,1,1,1
     ),col.main="Gray66")
      plot(v)
     #legend(.05, .9, legend = LABS[id], fill = COL[id],
     x="topleft",bty="n",)
     legend(.9, .9, legend = leg.txt, fill =
     COL[id],x="topleft",bty="n",text.col="Gray66",horiz=FALSE,b
     order=COL[id],cex=0.85)
     title(main="Native chondrocytes to monolayer chondrocytes")


#########################################
###Prioritised gene lists#################
#########################################

library(RobustRankAggreg)

     illumina_rank<-
     as.list(read.csv("illumina.csv",sep=",",header=TRUE))
     affymetrix_rank<-
     as.list(read.csv("affymetrix.csv",sep=",",header=TRUE))

     ill<-as.list(illumina_rank$Symbol.x)
     affy<-as.list(affymetrix_rank$Symbol.x)

     ill<-unlist(ill)
     affy<-unlist(affy)
     ill_rank<-as.character(ill)
     affy_rank<-as.character(affy)

     glist<-list(ill_rank,affy_rank)
     r = rankMatrix(glist)
     AGGREGATE<-aggregateRanks(glist,full=TRUE)
     head(AGGREGATE)
     write.csv(AGGREGATE,file="ranked_illumina_affymetrix.csv")

#####################################################################
```

4: Weighted gene co-expression network analysis of cartilage and tendon gene expression data

Abstract

The construction, and comparison, of gene-gene co-expression networks is a fundamental tool in systems biology and has the potential to extract biologically relevant gene regulatory sub-networks (modules) with strong phenotypic associations. Application of this methodology can facilitate the rational identification of central regulators, 'hubs', against which gain- or loss- of function studies can be designed.

This study outlines and applies a conceptual and methodological framework, weighted gene network co-expression analysis (WGCNA), to elucidate the regulatory sub-networks and hub genes that control differentiation status in chondrocytes and tenocytes in three-dimensional cultures. This approach was applied to two independent gene expression data sets from different microarray platforms and the global transcriptome network structures were compared. Defining modules with a strong phenotypic association with three-dimensional cultures may describe emergent behavior of organotypic systems with view to improving *in vitro* biomimetic cultures.

Consensus network analysis defined a preserved module of genes containing chondrocyte-associated genes *Pi15*, *Gpnmb* and *Serpina3n* identified as potential regulators of the alginate bead culture phenotype. The leucine zipper tumour suppressor (*Lstz2*) was identified as a potential modulator of dedifferentiation in a

monolayer-associated gene module. A siRNA study considering the effects of *Lzts2* knock-down on the expression of markers of differentiated status in chondrocytes produced equivocal results.

Additionally, this study demonstrated that small gene expression studies failed to robustly meet the criteria of scale-free topology required for this methodology, however, it represented a useful methodology for dealing with expression data from diverse sources and defined candidate regulators of the alginate bead culture phenotype.

4.1: Introduction

4.1.1: Dimensionality in gene expression data

A problem inherent to many microarray-based studies is the generation of highly dimensional data arising from tens of thousands of genes and relatively few biological samples (Wang, Miller et al. 2008). In the main the objective of these studies is to identify differentially expressed genes, which requires arbitrary thresholds to be set, often with the purpose of classifying groups. This 'hard-thresholding' results in prioritized gene lists representing the most extreme responses between two systems with no reference to the relationship between the genes (Conesa and Mortazavi 2014, Gaiteri, Ding et al. 2014). The dimensionality of these expression profiles may be reduced through the use of principal component analysis, supervised and unsupervised clustering techniques (Slonim 2002), as employed in the initial analysis of chondrocyte and tenocytes described in **Chapters 2** and **3**.

These methods, however, do not systematically study the interconnectivity of the individual genes identified, consequently they ignore the cumulative behavior of the regulatory pathways that contribute to perturb the gene expression patterns observed in differential gene expression analysis (Zhao, Langfelder et al. 2010). Methods that attempt to infer gene regulatory networks include Bayesian techniques and weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath 2008). These methods seek to describe the functions of the system rather than deconstructing and itemizing the component parts.

4.1.2: Conceptual understanding of gene co-expression networks and WGCNA methodology

Weighted gene co-expression network analysis (WGCNA) is a systems biology methodology that facilitates investigation of the global network properties of a transcriptome and provides functional insights into the organization of the network. The practical output of this type of analysis is the elucidation of higher-order relationships between groups of highly co-expressed genes (modules) and their associations with phenotypic traits. These modules represent the core units of the transcriptome network. It is possible, therefore, to also compare these functional units and the global network structure between gene expression studies or even across species (Miller, Horvath et al. 2010). Identification of the preservation of highly connected hubs within functional units across conditions can direct researchers to regulatory elements with potential phenotype-modulating properties.

Conceptually a co-expression network is relatively straightforward, although mathematically complex. Useful analogies are often made to social networks, which follow comparable network structures. Nodes, some connection point, represent genes that are expressed in a sample. Edges, or vertices, connect nodes based upon their co-expression across samples. WGCNA assumes that all nodes are connected and the connectivity has different strengths. The strength of the connections between nodes, in this case, indicates the importance of the genes in the networks and the measures are derived from the correlation of gene expression (Zhao, Langfelder et al. 2010).

These gene co-expression networks allow the exploration of the system level features and functionality of genes. As the co-expression of genes encodes the

downstream protein interactions the study of transcriptional co-expression patterns can reveal emergent properties of a cellular system. This application is a valid systems biology approach (Conesa and Mortazavi 2014) using comparisons of global transcriptomic network changes to define changes in the system. Network-associated methods are not limited to gene co-expression, but are widely applicable to biological networks (protein-protein interactions), as well as technological (e.g. world-wide web) and social interaction networks (Barabási and Oltvai 2004, Zhang and Horvath 2005). Global co-expression networks arise from a multitude of regulatory systems and information from every gene profiled may be used.

The network topology, the spatial relationships of the genes within the network, may have functional relevance. Networks may be reduced and binned into modules of highly correlated gene sets, which have been demonstrated to have functional commonality; these modules are not isolated and autonomous, but highly interconnected (Gaiteri, Ding et al. 2014).

The goals and philosophies of systems biology are well-placed to inform tissue engineering and regenerative medicine through the modeling of cell signalling and behavioural phenotypes (Cosgrove, Griffith et al. 2008) especially as regenerative strategies often aim to recapitulate dynamic processes, e.g. tissue morphogenesis. It is proposed that systems biology approaches (including co-expression network analysis) using multi-level data sources can facilitate the development of predictive models for bioengineered tissues to inform and refine these systems (Rajagopalan, Kasif et al. 2013). Systems biology tools are applied in this thesis to explore novel approaches to inform the future development of organotypic cultures.

Scale-free networks

Most biological networks are 'scale-free' (Barabási and Oltvai 2004, Albert 2005, Zhu, Gerstein et al. 2007), as opposed to random or hierarchical networks. Characteristically these scale-free networks follow a 'power-law' distribution, see below for clarification. In a scale-free network the probability that a node is connected to another node is statistically higher than in a random network; the properties of a scale free network are often defined by a small number of nodes that are highly connected - these may be referred to as 'hubs'.

Weighted gene co-expression network analysis (WGCNA) is based on the concept of 'scale-free topology' (Zhao, Langfelder et al. 2010), an overview of the methodology is presented in **Figure 4.1**. In a network that is considered to be 'scale-free' the connectivity, or degree, (k) of its components (nodes) follows a power-law distribution, such that:

$$p(k) \sim k^{-\gamma} \qquad \text{Equation 1.}$$

, where γ is the degree exponent. In other terms, the frequency distribution of degree, p(k), is the probability of a node having k links to other nodes of a network decays as a power-law (Barabási and Oltvai 2004). This confers a number of properties to scale-free networks:

i) The degree distribution decays as a function of the scaling parameter, the degree exponent, γ (above), such that the frequency distribution of node connectivity has a heavy tail;
ii) Scale-free networks are not homogenous with the network topology being defined by a few nodes of high degree or connectivity, which are linked to

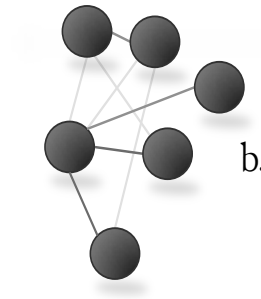
most other nodes in the network; by analogy, the Google search page or an international airport;

iii) There is considerable redundancy in a scale-free network and, therefore, a high tolerance of errors. The potential to communicate is not diminished by high failure rates;

iv) Scale-free networks are vulnerable to targeted attacks, i.e. the selection and removal of a hub node could negate communication throughout a system;

v) Growth and preferential attachment occur to already highly connected nodes, i.e. 'the rich get richer'.

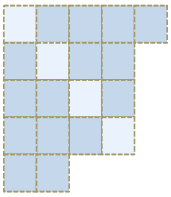
Figure 4.1: Overview of general approach to weighted gene co-expression network analysis – Results sections **4.3.1 – 4.3.3**



a.

Input is gene expression data. Nodes, balls (**a**), represent genes. The edges connecting them are associated with the absolute Pearson correlation coefficients as a measure of co-expression.

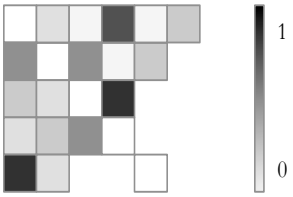
b.



An $n \times n$ similarity matrix (**b**), $S=[s_{ij}]$, describes the pairwise co-expression relationships between the genes i and j, such that: $s_{ij} = |cor(i,j)|$.

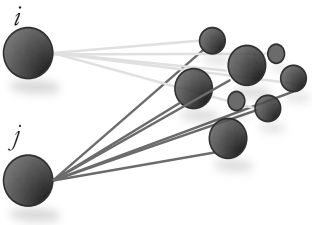
Gene expression data

c.



Using an adjacency function the similarity matrix is converted to a symmetrical adjacency matrix (**c**), $A=[a_{ij}]$ which encodes the connection strength between pairs of nodes, such that $a_{ij} \in [0,1]$. Diagonal elements of A are 1, i.e a_{ii}. This matrix is used to define node connectivity.

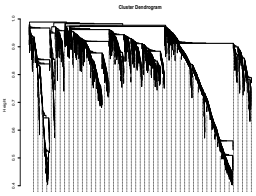
d.



Measures of node dissimilarity are defined from a topological overlap matrix (TOM), 1-TOM. Pairs of genes with high topological overlap (**d**) are likely to be functionally related.

Inputs

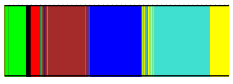
e.



Average linkage hierarchical clustering (**e**) methods used on the dissimilarity measures to detect modules.

h.

f.

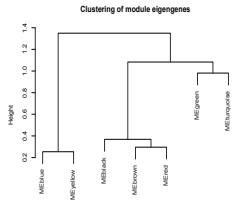


Gene modules (**f**), defined by colour code, are related to the 'branches' of the dendrogram as defined by a threshold cut height.

Quantitative data or phenotypic traits

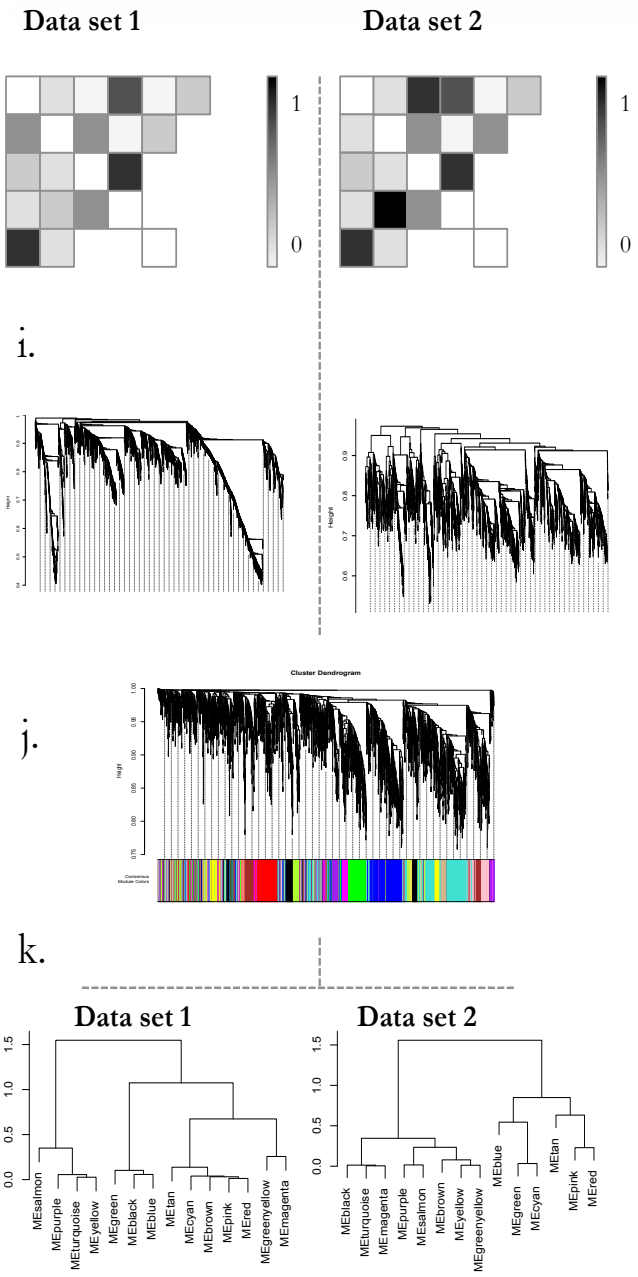
Continuous or categorical data (**h**) e.g. weight, bone mineral density, phenotypes, may be correlated with particular module eigengene.

g.



Networks are reduced to module representatives (**g**), module eigengenes, the correlations between which are used to form an eigengene network. This may reveal higher order meta-module organisation.

Figure 4.1 continued: Finding consensus modules across two data sets – Results section **4.3.4 – 4.3.6** .



Module and eigengene preservation statistics | meta-module identification | functional annotation | module-trait associations | hub-gene identification

Exploration of modules, eigengene meta-modules, and module-trait associations (**l**) to define biologically meaningful differences and similarities between data sets

Networks are constructed for each data set to consider the interaction patterns between genes (**i**). Data sets may be two comparable sources, species, reference vs. perturbed, etc.

Identify consensus modules (**j**), by consensus dissimilarity clustering, and define overlapping and preserved modules. Consensus modules may represent shared biological pathways differences in which may indicate alterations in regulation.

Using the eigengenes as the module representatives an **eigengene network**, a signed weighted gene co-expression network, is created (**k**) for each data set. This is based on the correlations between the eigengene for each module in a data set. This allows differential analysis of the network structure to take place by quantifying the preservation of consensus module relationships across two data sets. Preservation, or divergence, between the data sets may reveal changes in the biological pathways involved.

In order to construct a co-expression network complete gene expression data from microarray analysis is the input and the co-expression measure between genes is often defined by the absolute Pearson correlation coefficient, **Figure 4.1a** and **b**. Critically, to define that a connection, or co-expression edge, exists between two genes the Pearson correlation coefficient has to have a threshold; a 'hard' threshold would represent an absolute value of statistical significance (analogous to the arbitrary definitions of differential gene expression), however, defining connectivity in a dichomtomised manner (1, connected; 0, unconnected), is unlikely to be biologically relevant and would result in a considerable loss of information. In the general framework for the methodology 'soft'-thresholding is used to 'weight' each pairwise gene connection (Zhang and Horvath 2005) thereby encoding the relative importance of each gene.

An adjacency function is used to convert the matrix of co-expression similarity, the level of concordance in expression profiles across samples, into an adjacency matrix, which defines the connection 'strength' between each node/gene pair, *i* and *j* (Horvath 2011), **Figure 4.1c**. In other terms, an adjacency function allows the conversion of the original network into an alternative form. This adjacency matrix may be weighted through the application of a 'power adjacency function', β, (Equation 2) (Zhang and Horvath 2005).

$$a_{ij=}|cor(x_i,x_j)|^{\beta} \qquad \text{Equation 2}$$

The application of the power adjacency function changes the topological properties of a weighted network (Horvath 2011). A primary objective of co-expression network analysis is the detection of aggregates of nodes, termed 'modules', that are highly connected to each other. Methodologies used to define

these modules differ between techniques so only the approach used in this study is described.

Topological overlap and dissimilarity

A topological overlap matrix (TOM) is a similarity measure of relative connectivity formed from the adjacency matrix; 1-TOM defines the dissimilarity measure, which is the fundamental input to the clustering methods used to detect modules. An adjacency matrix represents each pair of genes a_{ij} in isolation, whilst a topological overlap matrix considers each pair of genes relative to all other genes within the network. Sets of genes have high topological overlap if they connect to approximately the same groups, or neighbourhood, of genes in the network, **Figure 4.1d**. Topological overlap is calculated on the comparative connectivity of a pair of genes to all other genes in a network (Yip and Horvath 2007). Two elements that have high topological overlap are more likely to have the same functionality than elements with lower topological overlap. Extended to genes it can be intimated that genes with high topological overlap are likely to be functionally comparable. In summary, this measure not only considers the expression correlation between two genes, but also, using a social networking analogy, how many 'friends' they share.

Identification of gene modules

Cell biology is inherently modular (Hartwell, Hopfield et al. 1999). Reductionist approaches attempted to reduce the understanding of biological phenomena to linear relationships. The ability, however, to apply discrete biological functions to individual molecules is rare; biological functions arise from molecular interactions organized into functional modules. Supervised and unsupervised clustering methods are widely used in genomic studies (e.g. hierarchical and *k*-means

clustering), each with their own limitations. The methodology applied here uses a dynamic hybrid approach where, first, hierarchical clustering is used to define a cluster dendrogram/tree, **Figure 4.1e and 4.1f**, with clusters defined as 'branches', then secondly uses a PAM-like (partition-around medioids) algorithm to define the clusters (Horvath 2011). Here, modules (used in place of the term 'clusters') are defined as genes that are highly connected to one another, and specifically within this methodology, genes with high topological overlap. These modules represent the fundamental functional unit of the transcriptional network (Miller, Horvath et al. 2010). As clustering techniques are non-robust (dendrograms can appear very differently depending on the form of hierarchical clustering used) the reproducibility of clusters/modules must be validated against other data sets. In this chapter modules defined in two independent data sets are used and validation techniques applied to assess their reproducibility. This approach is extended in **Chapter 5** to a meta-analysis across species.

Consensus networks and differential eigengene network analysis

In the context of systems biology functional modules span the knowledge gap between individual genes and the global properties emerging from the system (Zhao, Langfelder et al. 2010). As described above these co-expression modules represent the basic components of the system. When considered together co-expression modules may form a meta-network from which higher order organization of the transcriptome may be apparent. Module representatives, eigengenes, can be used to describe these meta-networks and are called eiegengene networks, **Figure 4.1k** and **4.1l**. Single eigengene network analysis may be used to describe the module relationships in a single data set, or differential eigengene network analysis, when used to compare the network relationships between data

sets. To achieve the latter WGCNA employs methods to define consensus modules from consensus dissimilarity measures derived from comparisons of the topological overlap matrices from the different data sets, **Figure 4.1j**. Whole network preservation between a reference and test data set may then be assessed using permutation tests. This study utilizes these methods to define co-expression modules conserved across two gene expression data sets profiling chondrocyte and tenocyte transcriptomes in different environmental conditions.

4.1.3: WGCNA: Recent applications

Weighted gene co-expression network analysis is now a well-established methodology that has been applied to diverse contexts. Recent work has demonstrated genes are highly connected in modules conserved across different types of cancer revealing robust prognostic signatures (Yang, Han et al. 2014). Horvath's group, in a seminal paper, demonstrated divergent co-expression modules in human and mouse brain transcriptomes with application to a better understanding of the relevance of murine models of human brain disorders (Miller, Horvath et al. 2010). Later work, contributing to an atlas of the brain transcriptome, showed that the topographical anatomy of the brain was reflected in the molecular topography (Hawrylycz, Lein et al. 2012). To date the author is only aware of two studies that apply this methodology to musculoskeletal tissues. In a study of chondrocyte differentiation Suwanwela and colleagues (2011) integrated co-expression network analysis of gene expression data from 27 mouse strains with quantitative data on bone geometry and bone mineral density (Suwanwela, Farber et al. 2011). In another study associating mouse gene expression data with bone mineral density Calabrese, *et al* (2012) identified a module with strong osteoblast-association; siRNA knock-down of an intra-

modular hub supported a role in osteoblasts proliferation and differentiation (Calabrese, Bennett et al. 2012). There are no publications considering cartilage and tendon gene expression responses to culture systems.

The rationale for the use of this type of technique is that it should permit the discovery of biologically interesting modules of genes with shared functionality in a phenotype-specific manner. Module preservation can be studied across different data sets to identify how conserved modules and their hubs are; it is suggested that this type of methodology may be more useful and robust than traditional meta-analysis techniques (Langfelder, Mischel et al. 2013).

Study hypothesis

It was hypothesized that emergent properties of three-dimensional culture systems would become apparent through the integration of co-expression network data from two gene expression data sets. By associating intra-modular hubs with cell-specific culture phenotypes gene targets may be identified that could inform improvements in organo-typic systems. It is proposed that the knock-down of gene transcripts for hubs with strong phenotypic associations for dedifferentiation could result in improved expression of markers of differentiated status. More broadly, the comparison of whole tissue and matrix-depleted chondrocyte and tenocyte transcriptome networks could define regulatory sub-networks relevant to the understanding of musculoskeletal degenerative disease.

4.2: Methods

4.2.1: Weighted Gene Co-expression Network Analysis

Data pre-processing, network construction and module detection

All analysis was undertaken in *R* using a comprehensive suite of functions implemented in the `WGCNA` (v1.41.1) package (Langfelder and Horvath 2008, Langfelder and Horvath 2012), **Figure 4.2a**. Equivalent codes for this methodology are provided in **Appendix 5.** The complete normalised expression data sets from Illumina (**Chapter 2**, n=36) and Affymetrix (**Chapter 3**, n=24) were filtered initially on variance and invariant genes were removed. In the preliminary analysis each data set was studied autonomously, **Figure 4.2b**. The expression preservation, the Pearson correlation of the ranked average gene expression, was calculated across these two data sets to assess the comparability of analysis across platforms.

The general co-expression network analysis methodology described by Zhang and Horvath (2005) (Zhang and Horvath 2005) was applied to these data sets. The soft threshold value, β, was chosen such that the lowest power that maintained an approximate scale-free topology was used.

The filtered genes were ranked on their connectivity and the top 3600 genes from each data set were retained for further analysis. A measure of node similarity, topological overlap, was calculated (Ravasz, Somera et al. 2002, Zhang and Horvath 2005). The dissimilarity measure, 1-TOM, was used as the input for average linkage hierarchical clustering to define gene modules. Modules were

merged at a dendrogram cut-height of 0.2 using a dynamic tree-cutting algorithm (Langfelder, Zhang et al. 2008), corresponding to a correlation of 1-0.2 = 0.80. All genes associated with a module were used as an input for gene ontology functional annotation using the *R* package `GOstats`, as before, and ensuring that the background for hypergeometric testing was restricted to the appropriate data source; ontology terms were statistically significant after FDR adjustments at $p<0.001$.

Module eigengenes were calculate to provide single representative expression pattern for a module (Langfelder and Horvath 2007). Correlations of eigengenes were used to form networks; these were plotted and manually assessed for correlation and the dendrogram cut height amended where necessary to merge highly correlated modules. The minimum entry to a module was retained at default settings to allow greater flexibility in later filtering for module hubs.

Consensus network and module preservation across data sets

To determine whether the set of modules identified in one data set were present in another a standard marginal model analysis was undertaken. A consensus network was derived from the Illumina and Affymetrix data sets and is defined as a single network created from the weighted average of correlation matrices from the separate data sets (Langfelder and Horvath 2007, Miller, Horvath et al. 2010). The consensus network was based upon the intersection of common Entrez identifiers across the two data sets after filtering of invariant data; this left 2795 genes, **Figure 4.2c**.

Module overlap, the number of genes common to two separate modules was used to defined how well modules identified in autonomous data set analysis were

conserved in consensus modules. This was further characterised by calculating the module preservation summary z-score from a series of permutation tests (Langfelder and Horvath 2007, Langfelder and Horvath 2008). Differential analysis of the eigengene networks were undertaken to define preservation of the network structures across data sets.

Consensus hub genes

Module Membership (kME) is the Pearson correlation between the expression level of each gene in a dataset and each module eigengene in a network. It quantifies the 'belonging' of a gene to a module. To define consensus module hub genes kME was calculated for all genes with module assignments and values ranked for each module; the genes with the highest kME in both networks were defined as consensus hubs.

Correlating modules to cellular conditions/traits

To determine whether modules were associated with the sample phenotypes the module eigengenes were correlated with a binary matrix defining phenotypic group membership. These phenotype, or trait, groups were inclusive of all samples with a common origin or condition, for example, "cartilage", "tendon", or "native".

Glossary

The unique terminology used within the WGCNA methodology is defined for reference, with respect to (Miller, Horvath et al. 2010). Terms associated with module analysis: a) Module eigengene – the first principal component of a module (**Figure 4.1g**); this represents a summary of the expression pattern characteristics of a module; b) module overlap – the number of common genes between two

modules arising from different data sets; c) module preservation – a collection of statistical tests that quantify how well module characteristics in one network are replicated in a second network.

Terms associated with correlation analysis are defined here: a) Gene Significance (GS), the absolute correlation between a module eigengene and a trait; b) the connections of a node within a module is the 'intramodular connectivity' defined as the summed connections of a node with all other nodes within a module; c) a global module eigengene-based connectivity measure, Module Membership (*k*ME), defined above.

4.2.2: Data visualization and network representation

Modules were represented graphically using Cytoscape (v3.1.1, October 2014), (Cline, Smoot et al. 2007, Killcoyne, Carter et al. 2009). The network structure, consisting of nodes (genes filtered for high module membership) and edges (weighted intra-modular connections based upon the topological overlap matrix) were exported to Cytoscape. The web-application STRING, version 9.1 (Franceschini, Szklarczyk et al. 2013), (http://string-db.org) was used to define protein-protein interactions between genes identified to be consensus hub genes. Circos plots were rendered using the Circos web application (http://circos.ca) (Krzywinski, Schein et al. 2009) to summarise module overlaps between data sets.

4.2.3: Silencing of *Lztś2* expression using siRNA

Chondrocytes were obtained from 12-week old male Wistar rats (n=3, **Chapter 2** stock), **Figure 4.2a**. Passage three chondrocytes cells were plated in 24-well culture plates at a density of 2.5 x10^4 cells/well and allowed to adhere overnight (~12 h) in complete media, free of phenol red. For each biological replicate

siRNA experiments were performed in triplicate and these were pooled prior to RNA extraction.

Prior to transfection all instrumentation, Eppendorfs, micro-pipettes and reagent containers were treated with RNAzap® (Ambion, Life Technologies) according to the manufacturers instructions. Optimal concentrations of siRNA and Lipofectamine® 2000 were derived from comparisons of maximal transfection efficiency based upon varying concentration of siRNA (15 or 30 ρmols), Lipofectamine® 2000 (0.5-4 mL), and cell number (2.5 $x10^4$ – $x10^4$) per well in a preliminary study.

Small interfering RNAs (siRNAs) directed against the leucine zipper, putative tumour suppressor 2, *Lzts2*, transcripts (Rat Silencer Select siRNA (s1722155, #4390771) (Ambion, Life Technologies)) were used at 30 ρmols per well. Equimolar concentrations of negative control (Silencer Select Negative Control No. 1 siRNA (AM4611). Transfection efficiency was assessed using a Cy3-labeled control (Silencer Cy3-Labelled Negative Control siRNA). Cells were transfected with siRNA incubated with 2.5 μl Lipofectamine® 2000 Transfection Reagent (Ambion, as before) according to the manufacturer's guidelines in a final volume of 0.5 mL of CM1, as before. Adherent cells were washed with warmed PBS twice to remove media and FBS. To each well 100 μl of the siRNA/lipofectamine solution was added; additionally 400 μl of Optim-MEM (with 3 % FBS, no antibiotics/anti-fungal) was added. Cells were incubated at 37 °C for 6 hours; after this time media was removed, cells washed with PBS and 500 μl of complete media added. Cells were left overnight (>12 hrs). Cells transfected with Cy3-labelled negative control siRNA were inspected under green fluorescent light (547-563 nm) for evidence of transfection. Cultures were harvested and stored in Trizol

at -80 °C for RNA extraction as previously described. Following RNA quality control RNA some samples were re-precipitated and further concentration steps performed. Briefly, sterile filtered 5M ammonium acetate (Ambion, Applied Biosystems) was added to the RNA suspension to bring the final concentration of ammonium acetate to 2.5M; 2 µl of glycogen/20 µl RNA solution was subsequently added. To this 2.5 volumes of 100 % molecular grade ethanol (EtOH) (Sigma-Aldrich) was added and the solution was frozen at -80 °C overnight. Following this samples were allowed to defrost on ice, were centrifuged at 20,000 * *g* for 15 minutes at 4 °C. The supernatant was discarded and the RNA pellet washed with 1 mL glacial 75 % EtOH to remove residual salt. RNA pellets were centrifuged as before, supernatant discarded once more and the residual EtOH allowed to evaporate. Pellets were re-suspended with 10 µl of RNAse-free water.

4.2.4: Quantitative PCR

The design and validation of primer pairs targeting *Lzts2* was undertaken as described in **Chapter 2**. The primer pairs were designed to target exon two of the *Lzts2* transcript as this was the siRNA target: 5'-GATCCCCGAGAACATCAGGC (60.25 °C, forward), and 5'-TCTCCATATTCTTCTCCAGCCTTC (59.65 °C, reverse), producing a 94 base product. Previously described qPCR protocols were applied to this study. In a serial ten-fold dilution study the primer pair had a 90 % amplification efficiency. Expression analysis was undertaken using the ΔC_t method, as before. Statistical analysis was undertaken on the $2^{\wedge -\Delta Ct}$ values by ANOVA (Tukey *post-hoc* analysis) after a Shapiro-Wilks test to assess deviations from a Gaussian distribution.

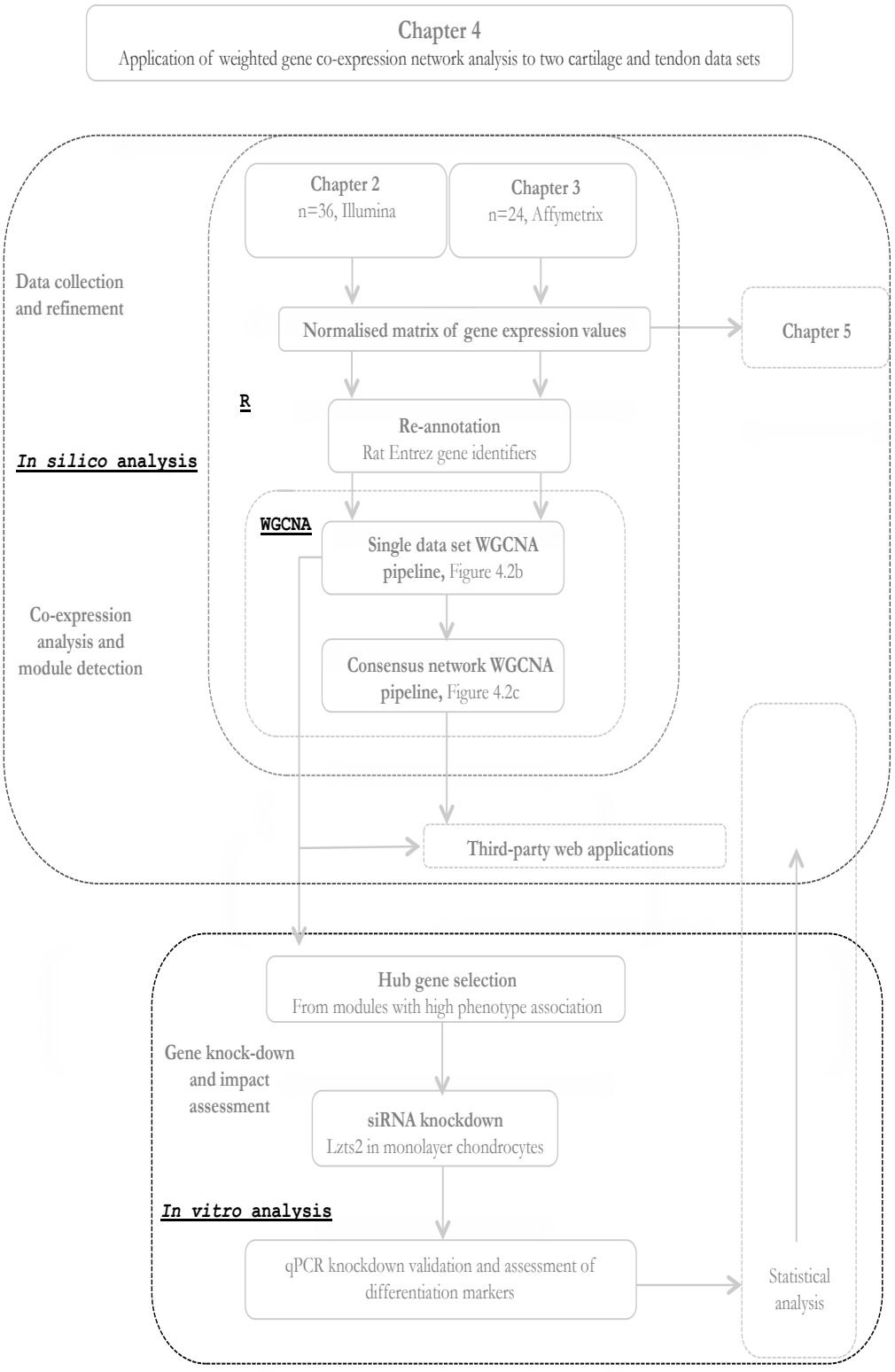


Figure 4.2a: Overview of experimental design for results presented in **Chapter 4**. Input data is derived from **Chapters 2** and **3** using Illumina and Affymetrix microarray gene expression data respectively. Each data set follows, i) a round of independent network generation and module detection and, ii) consensus network generation and comparative module analysis. Only hub gene detected in Illumina data is used for siRNA study. Specific data analysis pipeline follows that described in **Figure 4.2b** (single data set) and **c** (comparative analysis). Expression data also used in subsequent analysis in **Chapter 5**. Statistical analysis of qPCR data was undertaken in *R*.

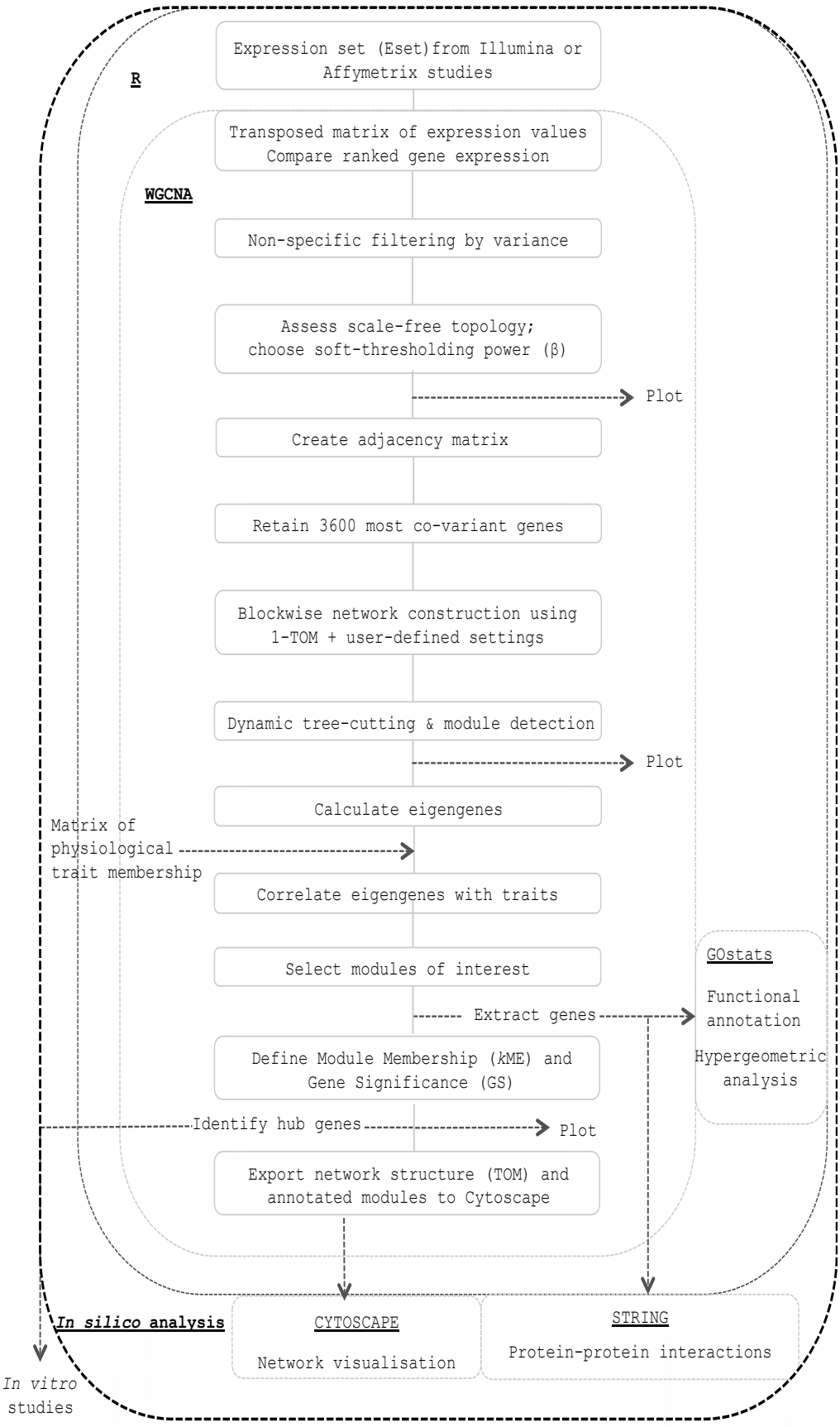


Figure 4.2b: Data analysis pipeline within *R* using a suite of functions within the WGCNA package. Pipeline represents the individual analysis undertaken for each data set. Specific functions are defined in **Appendix 4** *R* codes. Third-party web applications are indicated.

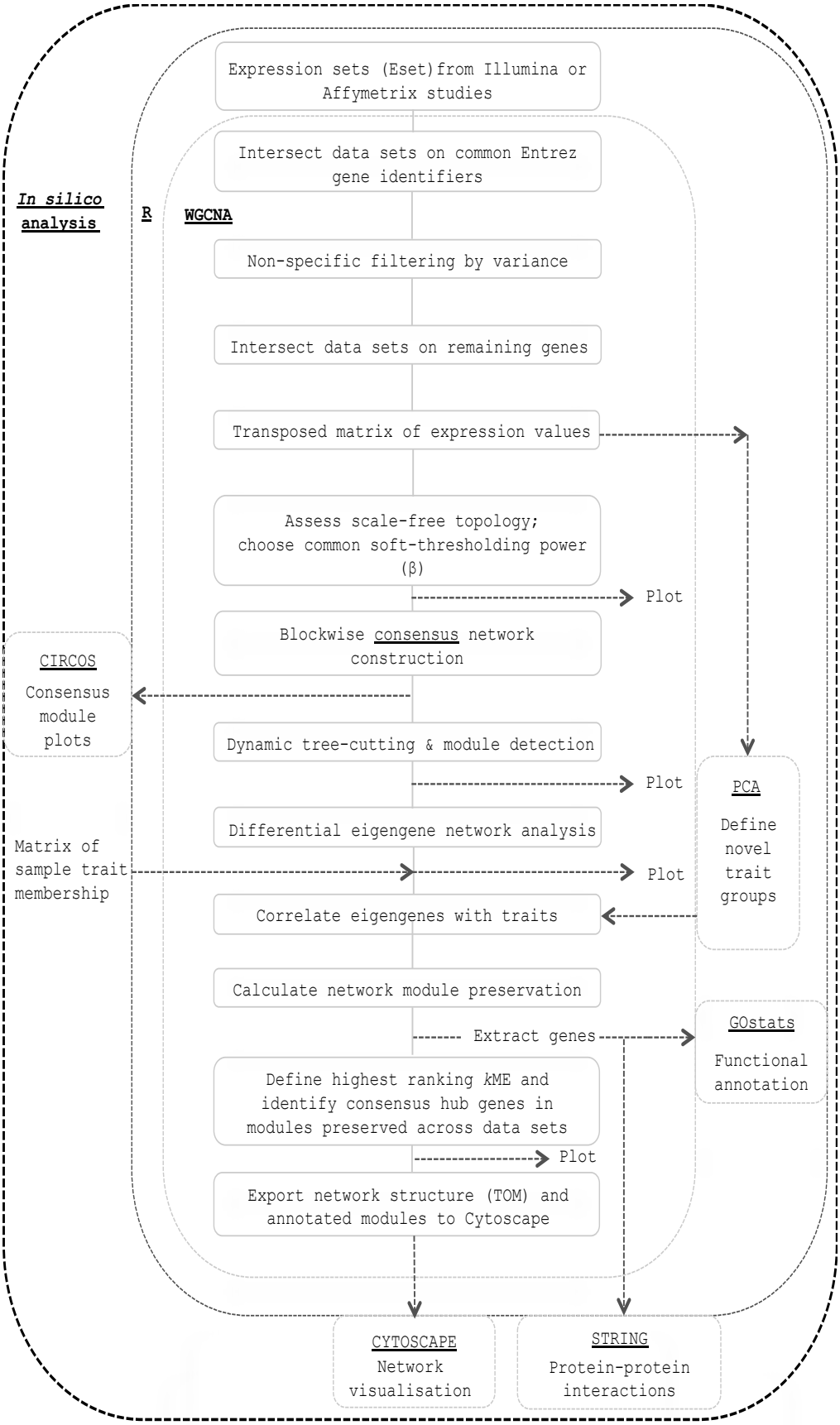


Figure 4.2c: Data analysis pipeline within *R* using a suite of functions within the WGCNA package. Pipeline represents the consensus analysis undertaken to compare the network structures of two independent data sets, e.g. conditions, species. Preserved (consensus) and unique modules were defined and annotated. Specific functions are defined in **Appendix 4** *R* codes. Third-party web applications are indicated.

4.3: Results

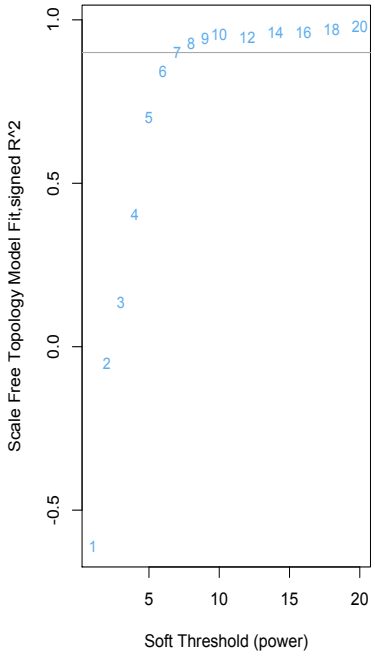
4.3.1: Data preparation, module formation and functional annotation

Network topology and module detection

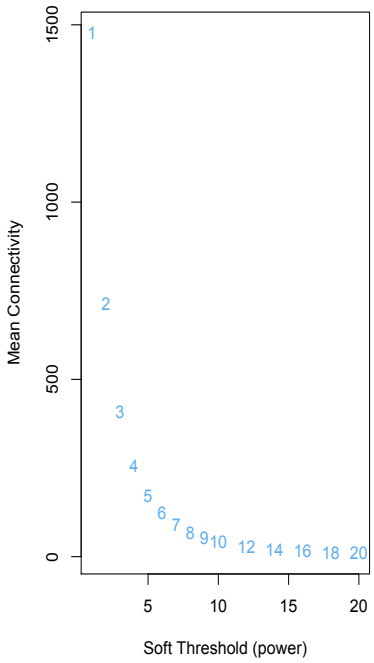
Initially the data was considered against the general scale-free topology criterion to define the adjacency function parameters. The soft-thresholding powers which resulted in an approximate scale-free topology for both Illumina ($\beta = 9$) and Affymetrix ($\beta = 12$) data was determined. The datasets differed in their network topology with the Affymetrix data only having approximate scale-free topology at higher soft-threshold values ($R^2 < 0.72$ at $\beta = 12$), **Figure 4.2** and **4.3**. This would normally not satisfy the scale-free topology criterion ($R^2 > 0.8$), however, in a plot of the regression line between $\log_{10}(p(k))$ and $\log_{10}(p(k))$ the slope was -1.13, comparable to that of the Illumina data, which did meet scale-free topology criterion.

Module detection using hierarchical clustering of the dissimilarity matrix defined seven module eigengenes for the Illumina data (**Figure 4.5**) and six for the Affymetrix (**Figure 4.6**) dataset. Genes included within a module were functionally annotated using gene ontology terms ($p<0.001$), **Tables 4.1** and **4.2**. Modules were allocated arbitrary colours for each dataset - these are not comparable between data sets. For clarity an alphanumeric code is employed as defined in the previous tables. Broadly, there were modules that shared functional annotations across data sets, in particular those relating to metabolic and cell cycle processes, but also immune and defense response annotations and anatomical structure development.

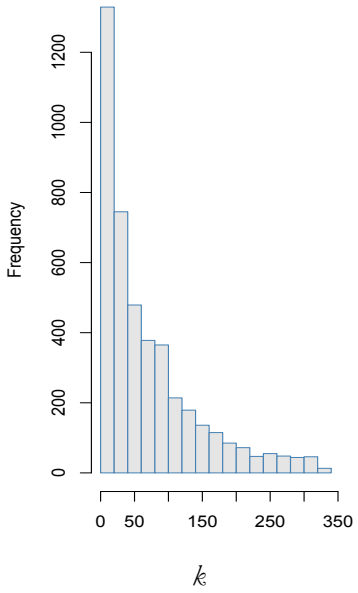
i. Scale independence



ii. Mean connectivity



iii. Histogram of connectivities, *k*



iv. Assessment of scale-free topology Scale, R^2=0.92, slope=-1.19

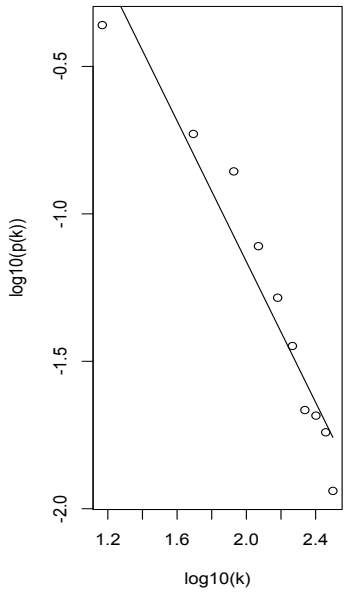


Figure 4.3: Upper panel - Assessment of **Illumina** data for weighted gene co-expression network analysis. **i**) Scale-free topology index (y-axis) as a function of the soft-thresholding power, β (x-axis), left panel. Intersection line at R^2=0.9. Right Panel **ii**) shows the mean connectivity, or degree, (y-axis) as a function of the soft-thresholding power (x-axis). Degree decreases with increasing soft-threshold value. Lower panel – (**iii**) histogram of connectivity values (*k*), (**iv**) log-log plot of the connectivities fitted to a linear model. The R^2 value, the square of the correlation between log10(p(*k*)) and log10(*k*), can be read as an index of the scale-freedom of the network topology. The Illumina data is shown to have approximate scale-free topology. Beta, β=8 is chosen as it is an effective trade-off between maximising scale-free topology and retaining a high mean number of connections.

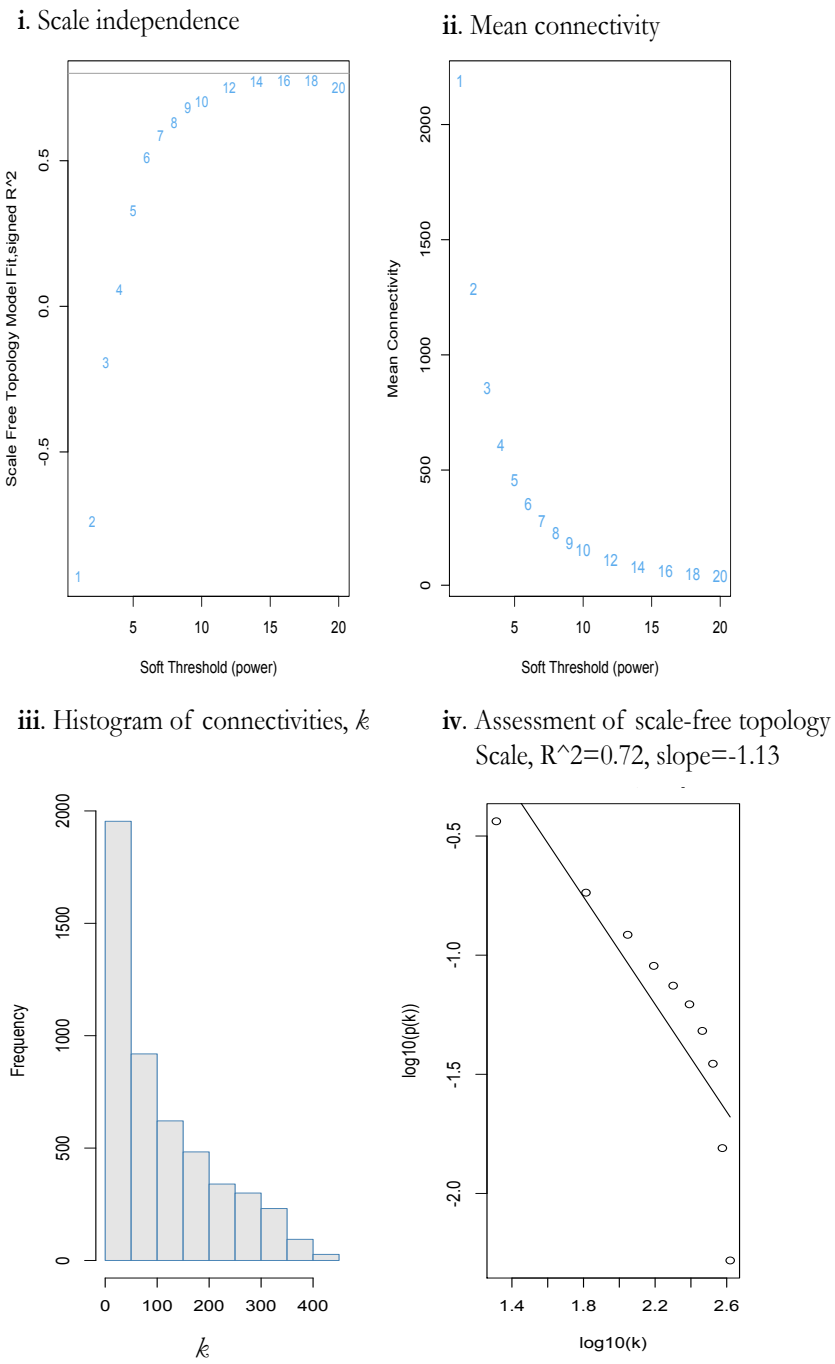


Figure 4.4: Upper panel - Assessment of **Affymetrix** data for weighted gene co-expression network analysis. **i**) Scale-free topology index (y-axis) as a function of the soft-thresholding power, β (x-axis), left panel. Intersection line at R^2=0.8. Right panel **ii**) shows the mean connectivity, or degree, (y-axis) as a function of the soft-thresholding power (x-axis). Lower panel – (**iii**) histogram of connectivity values (*k*), (**iv**) log-log plot of the connectivities fitted to a linear model. The Affymetrix data is shown to have only moderate scale-free topology. Beta, β=12 is chosen as it is an effective trade-off between maximising scale-free topology and retaining a high mean number of connections. Although R^2=0.8 is the recommended threshold value for approximate scale-free topology the slope of the linear regression between log10(p(*k*)) and log10(*k*) is <-1 and comparable to the Illumina data. On this basis the data was retained for further analysis.

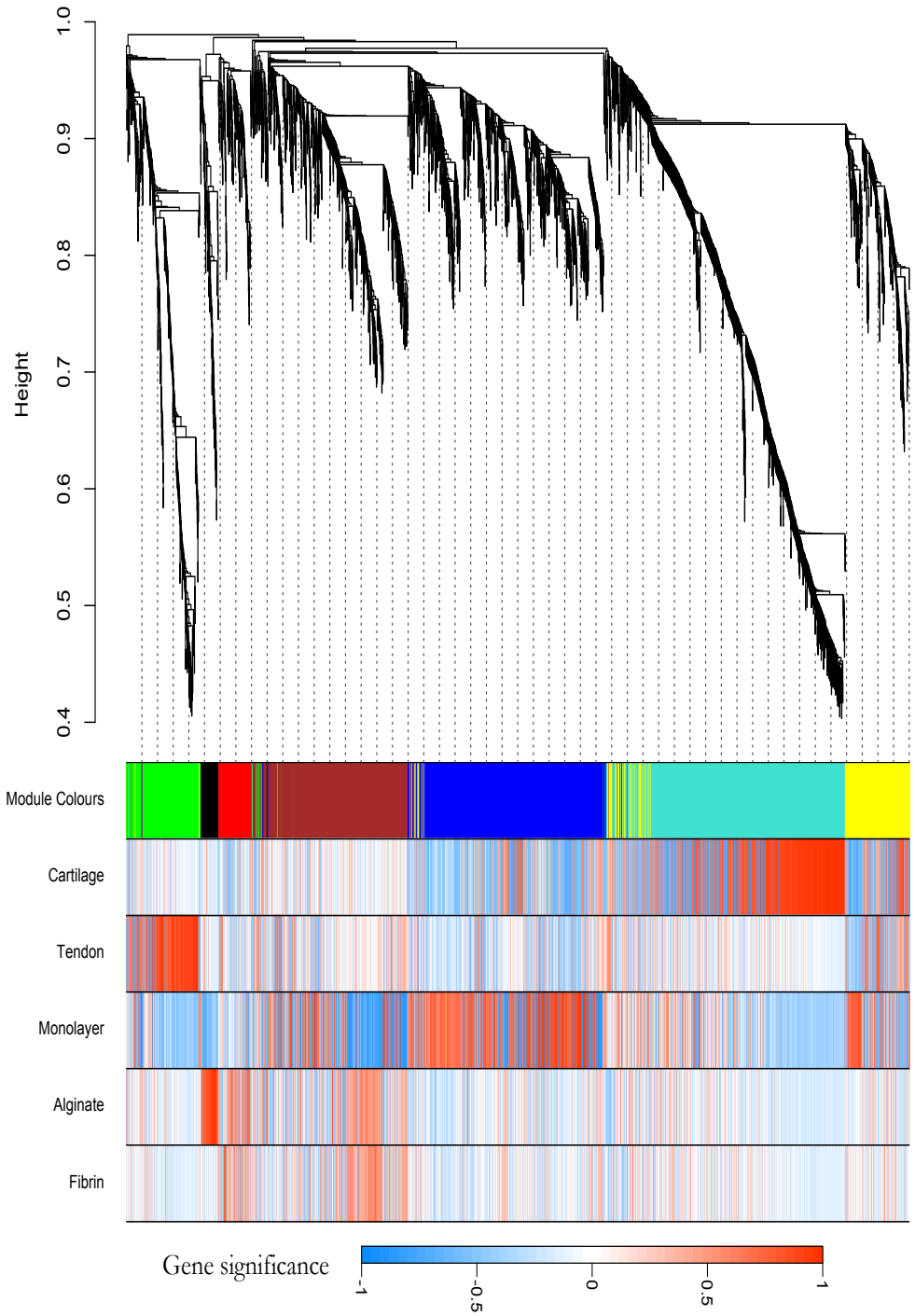


Figure 4.5: Gene co-expression dendrogram (top) and modules with phenotypic trait correlations (below) for **Illumina** data. Hierarchical clustering of genes in dendrogram is based upon 1-TOM (dissimilarity matrix) with modules (top row, coloured blocks) determined using a dynamic tree-cutting algorithm at 0.2. The five linear heatmap plots define the gene significance (GS) values – the absolute correlation between the gene and trait, heatmap legend. The heatmap plots demonstrate that the genes with the highest association for a trait (defined in **Table 4.1**) are located in the distal extremities of the dendrogram branches, for example the cartilage associated module for the Illumina data is the turquoise module; the genes with the highest gene significance are in the far right of the turquoise block and associated with the tips of the dendrogram branch.

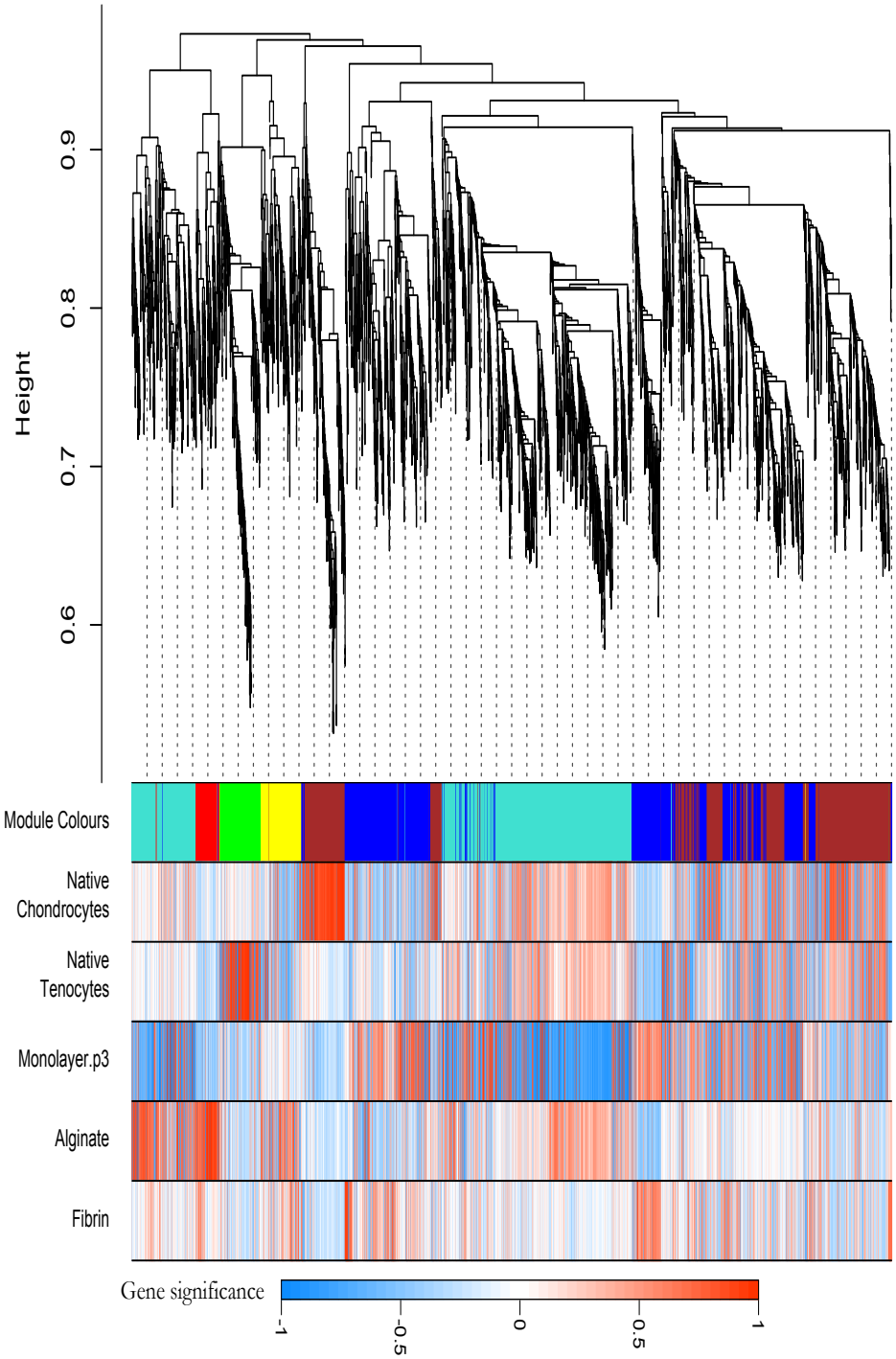


Figure 4.6: Gene co-expression dendrogram (top) and modules with phenotypic trait correlations (below) for **Affymetrix** data (n=24). Six modules were defined for the Affymetrix data on the top 3600 most connected genes. In contrast to the Illumina data modules are not as defined and module eigengenes more strongly correlated to each other as demonstrated by the height legend (y-axis). This dendrogram is based on a dissimilarity adjacency matrix and so the longer the branch the more dissimilar it is to other modules. This is also reflected in the gene significance distribution for different traits (linear heatmaps). Although some clear associations are present, e.g. native chondrocytes with the brown module, association is much more variable across phenotypes

Table 4.1: Gene ontology annotation for modules defined from **Illumina** data set. Each module has an arbitrary colour allocated, which is comparable only between other modules from the same data set. Gene ontology terms for biological process, metabolic function and cell compartment are significant at $p<0.001$ after hypergeometric testing and FDR adjustment of p-values; values are shown in parentheses. Universe for hypergeometric testing was genes from the Illumina RatRefv1 microarray with valid Entrez gene identifiers. Duplicates were removed. Terms provided are representative, but not exhaustive. Alphanumeric codes are used to unambiguously define modules between analysis and are used in the text.

Reference code for Illumina (ILL)-specific modules	Biological Process	Metabolic Function	Cell Compartment
ILL1	Response to organo-nitrogen compound (3.1e-9) Inflammatory response (1.3e-07) Response to cytokine stimulus (2.5e-5) NIK/NF-kappaB cascade (2.2e-4)	Methyl indole-3-acetate esterase activity (1.1e-5) Protein binding (2.1e-4) MRF binding (2.6e-4) Protein heterodimerisation activity (4.1e-4)	Extracellular space (1.6e-5)
ILL2	Cellular protein metabolic process (2.4e-11) Gene expression (2.4e-6) Cellular response to stress (3.8e-5) Apoptotic signalling pathway (4.9e-4)	Protein binding (5.1e-15) RNA binding (2.9e-5) Peptidase activator activity (3.03e-4)	Intracellular part (4.9e-30) Ribonucleoprotein complex (6.02e-9)
ILL3	Translation (5.3e-14) Metabolic process (7.7e-11) Organic substance metabolic process (4.6e-8) Nitrogen compound metabolic process (5.5e-4)	Structural constituent of ribosome (1.6e-15) RNA binding (3.2e-8) Translation factor activity, nucleic acid binding (3.2e-4)	Intracellular part (1.4e-35) Mitochondrion (6.8e-15)
ILL4	Muscle system process (2.9e-14) Skeletal muscle tissue development (1.6e-6) Actin filmament-based process (1.02e-5) Tissue regeneration (4.8e-5)	Structural constituent of muscle (9.1e-12) Cytoskeletal protein binding (9.8e-11) Actin binding (3.2e-7)	Myofibril (1.03e-27) Actin cytoskeleton (5.9e-11) Cytoskeleton (6.3e-6)
ILL5	Response to external stimulus (2.7e-4) Bone mineralisation (6.8e-4) Response to stress (6.8e-4) Biomineral tissue development (8.2e-4)	Proton-transporting ATPase activity (3.1e-4)	Cytoplasm (4.7e-6) Mitochondrion (5.9-6) Extracellular matrix (7.8e-4)
ILL6	Cell cycle (1.7e-10) Immune system process (6.3e-10) Response to biotic stimulus (6.5e-9) Co-factor biosynthetic process (5.7e-6)	Binding (1.9e-16) Catalytic activity (1.2e-11) Oxidoreductase activity (5.3e-5)	Intracellular part (3.2e-21) Nucleus (2.8e-13) Cytoplasm (4.8e-11)
ILL7	Cellular response to organic substance (8.7e-5) Cell adhesion (8.7e-5) Response to cytokine stimulus (8.8e-5) Anatomical structure development (1.7e-4)	Oligosaccharyl transferase activity (8.2e-4)	Cytoplasmic part (2.2e-12) Intracellular (9.5e-9) Lysosome (5.5e-8)

Table 4.2: Gene ontology annotation for modules defined from **Affymetrix** data set. Each module has an arbitrary colour allocated, which is comparable only between other modules from the same data set. Gene ontology terms for biological process, metabolic function and cell compartment are significant at $p<0.001$ after hypergeometric testing and FDR adjustment of p-values; values are shown in parentheses. Universe for hypergeometric testing was genes from the Affymetrix Gene ST 2.0 microarray with valid Entrez gene identifiers. Duplicates were removed. Alphanumeric codes are used to unambiguously define modules between analysis and are used in the text.

Reference code for Affymetrix (AFF)-specific modules	Biological Process	Metabolic Function	Cell Compartment
AFF1	Cellular macromolecule metabolic process (4.9e-20) Gene expression (2.2e-18) Biosynthetic process (2.4e-15)	Binding (2.6e-22) Organic cyclic compound binding (2.7e-15) DNA binding (3.1e-9)	Intracellular part (6.9e-41) Organelle (2.5e-30) Nucleus (3.9e-27)
AFF2	Immune system process (3.5e-13) Response to stress (6.7e-11) Cell activation (2e-8)	Protein binding (2.2e-10) PI3-kinase regulator activity (6.8e-4) Chemokine receptor binding (6.8e-4)	Cytoplasm (9.02e-9) Actin filament (8.9e-5) Intracellular (7.3e-4)
AFF3	Blood vessel morphogenesis (1.2e-7) Cardiovascular system development (2.8e-6) Anatomical structure formation involved in morphogenesis (1.1e-5)	Cyclase activity (5.5e-4) Guanylate cyclase activity (6.1e-4) Ion binding (8.5e-4)	Plasma membrane (1.2e-4) Cell periphery (1.4e-4) Membrane (1.7e-4)
AFF4	Mitotic cell cycle (2.2e-12) Developmental process (9.7e-9) Anatomical structure development (2.02e-7)	Protein binding (1.2e-17) Cytoskeletal protein binding (1.01e-11) ECM structural constituent (4.7e-5)	Cytoskeleton (1.5e-10) Extracellular matrix (4.6e-10) Stress fibre (5.8e-7)
AFF5	Response to external stimulus (2.9e-4) Regulation of signalling (2.9e-4) Negative regulation of MAPK cascade (6.9e-4)	Receptor binding (2.3e-4)	Extracellular region part (2.3e-4)
AFF6	Anatomical structure morphogenesis (5.5e-7) Developmental process (3.2e-6) Tissue morphogenesis (4.2e-5)	Carbohydrate binding (3.1e-4) Coreceptor activity (8.7e-4) Receptor tyrosine kinase binding (8.7e-4)	Membrane raft (5.3e-4) Plasma membrane part (6.7e-4) Vesicle (7.8e-4)

Relating modules to phenotypic traits

In order to evaluate whether any of the data set-specific modules had associations with the sample phenotypes a binary matrix was prepared to define membership of a phenotype, e.g. alginate cultures. These phenotypes, or traits, were correlated with the module eigengenes for each network. Illumina gene expression data showed divergent module eigengenes for cartilage and tendon, **Figure 4.7**. In line with the findings in **Chapter 3** co-expression module AFF2 for Affymetrix data was more highly correlated with both native matrix-depleted chondrocytes and tenocytes than either group alone, suggesting a closer phenotype, **Figure 4.8**. Alginate cultures in both datasets were found to have strong module eigengene associations (ILL1 and AFF5), but associations were equivocal for fibrin cultures alone.

Relating the module-trait associations back to the biological process functional annotations (**Table 4.1** and **4.2**) confirmed the annotations associated with the differential expression analysis in **Chapter 2** and **3**. For example, the alginate trait was association with the Illumina ILL1-module eigengene (cor = 0.97, p = 2e-21) and the terms 'inflammatory response' and 'response to cytokine stimulus', **Figure 4.7**. In the matrix-depleted chondrocytes from alginate beads in the Affymetrix data these samples were associated with the AFF5 module, **Figure 4.8**, (cor = 0.92, p = 2e-10) and the terms 'response to external stimulus' and 'negative regulation of MAPK cascade'.

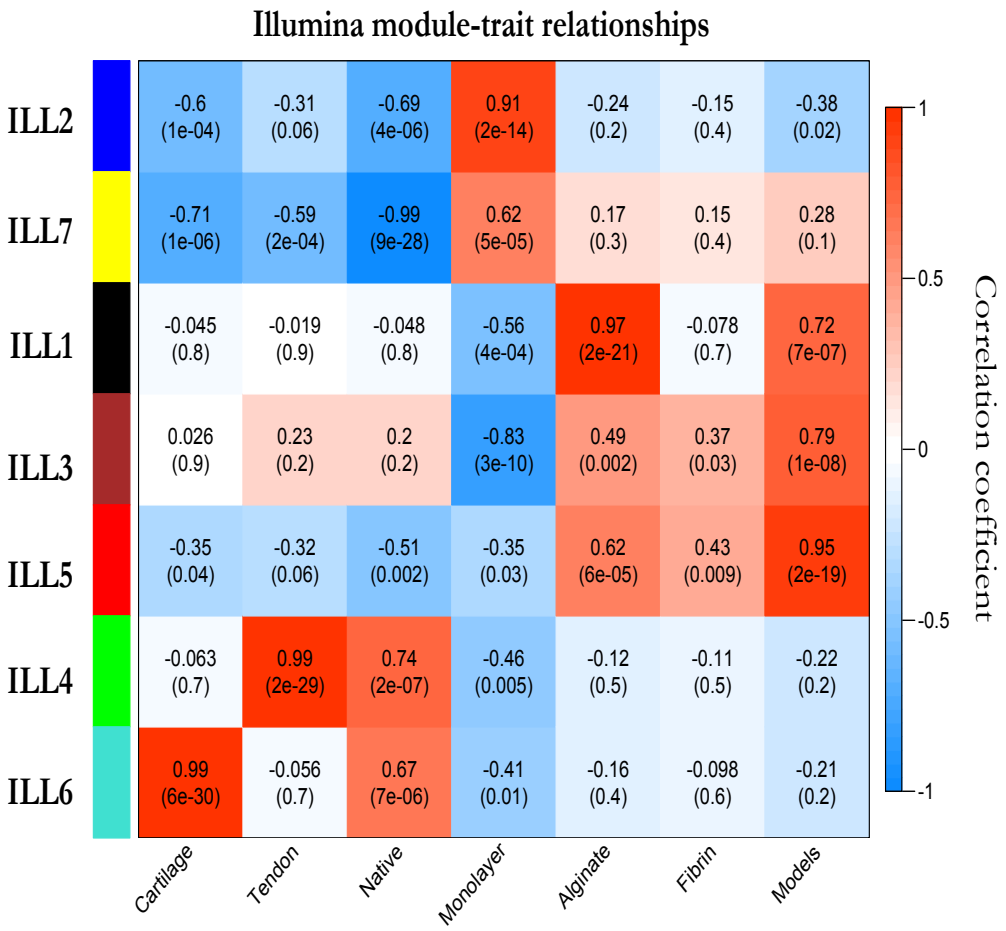


Figure 4.7: Associations between **Illumina** module eigengenes (rows) and sample traits (columns). Within each cell is the corresponding correlation (coded by colour – heatmap) and *p*-value (parenthesis). Samples may be members of more than one trait, for example, 'Native' comprises both cartilage and tendon samples, whilst 'Models' are an aggregation of both alginate and fibrin culture samples. Strong associations are found between the ILL6 and ILL4 modules with the native cartilage and tendon samples respectively. Monolayer (ILL2 module eigengene) and alginate (ILL1 module eigengene) traits also show strong associations. Weaker associations are found for the ILL5 and ILL3 module eigengenes and various traits.

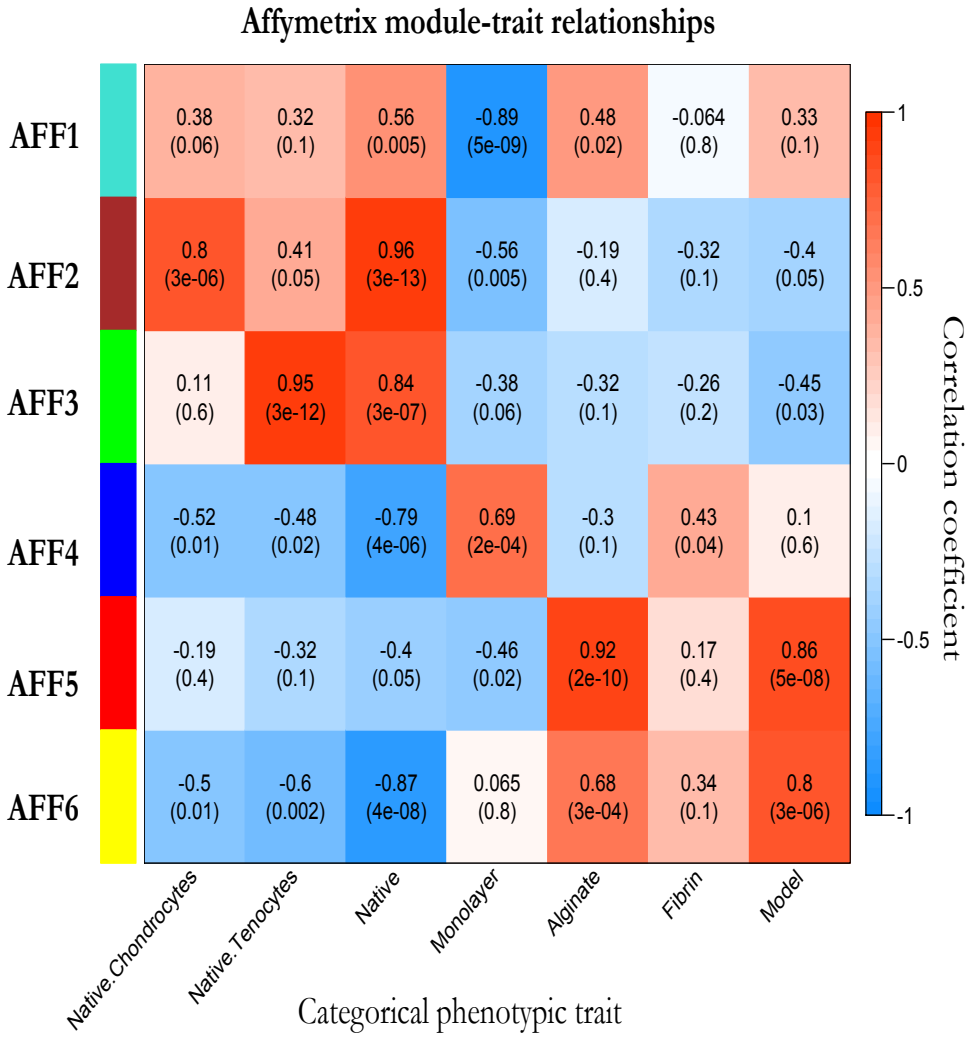


Figure 4.8: Associations between **Affymetrix** module eigengenes (rows) and sample traits (columns) as described above. Here matrix depleted cells were the input, native chondrocytes and tenocytes are cells isolated directly from tissue; monolayer is passage three. Strong associations are found between native samples and the AFF2 and AFF3 modules, whilst three-dimensional culture systems are more highly associated with the AFF5 and AFF6 modules. The monolayer trait appears associated with the AFF4 module, but this is not a strong statistical association.

4.3.2: Gene significance and module membership

In order to identify genes that are central elements to functional modules, i.e. module hubs, and also strongly associated with a phenotypic trait, three quantitative measures were calculated – Gene Significance (GS), intramodular connectivity, and Module Membership (*k*ME). Gene significance for every gene against each module is shown in **Figures 4.5 and 4.6**. As Module Membership is the more practical value for network comparisons this was used over intramodular connectivity, however, there was high correlation between the two values (data not shown).

For the Illumina data there was evidence to support the statement that genes with strong trait association were also the most highly connected within a module and had high Module Membership (correlation > 0.9), **Figure 4.9.** The cartilage- and tendon-associated modules (ILL6 and ILL4) had high correlation between GS and *k*ME (cartilage - 0.99, $p<1e\text{-}200$; tendon – 0.98, $p<1e\text{-}200$). This would indicate that genes that are highly associated with these phenotypes are also likely to be highly connected genes in these modules and could be considered hub elements. In contrast the ILL3 module, which had moderate association with three-dimensional culture models, demonstrated poor correlation between GS and *k*ME (cor = 0.34) indicating that the assumption that high trait association was related to high gene connectivity did not hold true in this module. In comparison, for the Affymetrix data similar confidence could not be extended to native cells, monolayer and alginate culture models and their associated modules, with correlations below 0.9 evident. Divergent relationships were evident for the combined native cell phenotype; this was consistent with two cell populations contributing to this analysis, **Figure 4.10**.

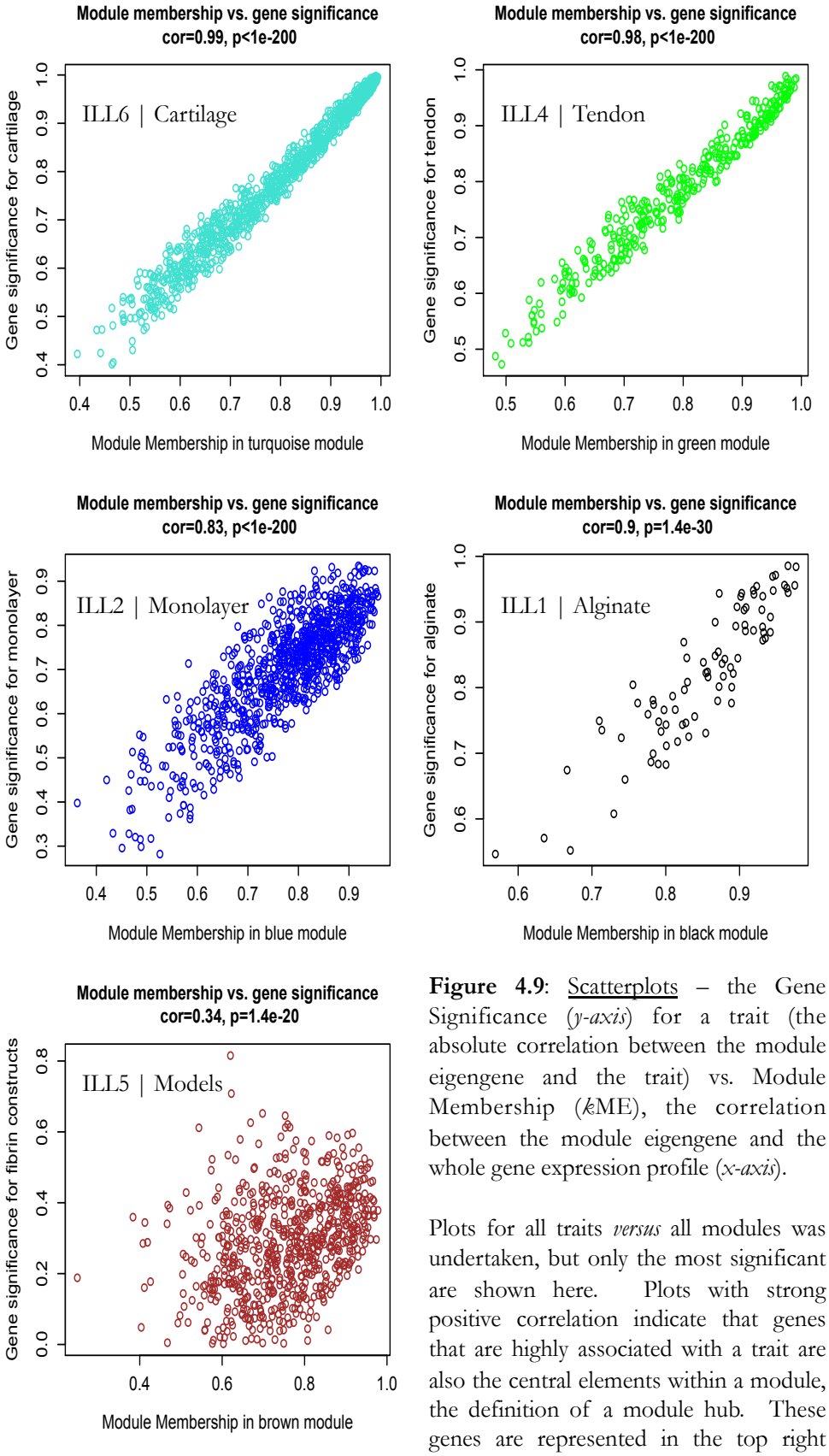


Figure 4.9: Scatterplots – the Gene Significance (*y-axis*) for a trait (the absolute correlation between the module eigengene and the trait) vs. Module Membership (*k*ME), the correlation between the module eigengene and the whole gene expression profile (*x-axis*).

Plots for all traits *versus* all modules was undertaken, but only the most significant are shown here. Plots with strong positive correlation indicate that genes that are highly associated with a trait are also the central elements within a module, the definition of a module hub. These genes are represented in the top right corner of each plot. This relationship is true for modules ILL1, ILL4 and ILL6, however, for low correlations (ILL5) hub genes cannot be confidently identified.

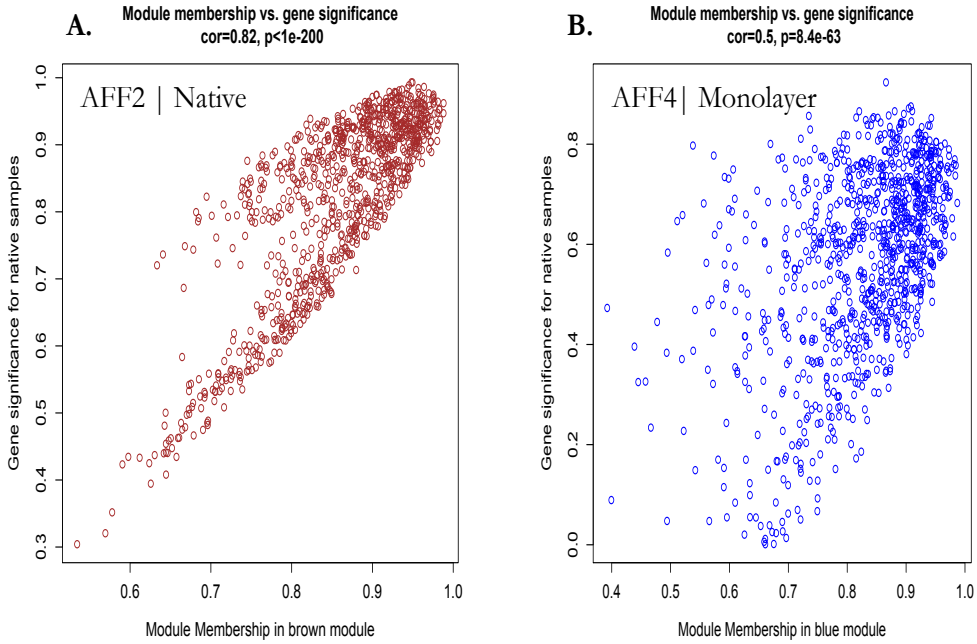


Figure 4.10: Scatterplot of Gene Significance (GS) (*y-axis*) versus Module Membership (*k*ME) (*x-axis*) relating to Affymetrix data. Associated modules and phenotypic traits are provided. Moderate correlations between GS for a trait and *k*ME were only found for the **A**: 'native' cell phenotype (matrix-depleted cells isolated from cartilage and tendon); **B**: monolayer cultures, and alginate beads (not shown). Although the AFF2 module eigengene had a high correlation with the matrix-depleted phenotype from native cells (cor = 0.96) the relationship between GS and *k*ME was only highly correlated at high values of GS and *k*ME with divergent relationships evident with genes that had lower associations with the phenotypic trait indicative of the presence of two phenotypes. The monolayer associated module demonstrated poor associations between GS and *k*ME indicating that hub genes for this module could not be confidently identified.

4.3.3: Defining putative module hubs

For modules with high trait significance, and for which there was a strong association between Gene Significance and Module Membership (*k*ME), the predicted hubs were chosen as the top twenty genes with a *k*ME >0.9, consistent with standard methodology. Gene ontology annotation was performed on the top 50 genes where *k*ME >0.9 and GS >0.5. Results for both network analysis are shown in **Table 4.3** and **4.4**.

Native chondrocyte-associated modules

The ILL6 module (1007 genes) had the highest association with the trait 'cartilage'. The genes with the highest *k*ME for 'cartilage' included the cathepsins *Ctse* and *Ctsg*, and genes identified as highly expressed in both cartilage datasets, *Sell* (selectin L) and *Dmp1* (dentin matrix acidic phosphoprotein 1). In the Affymetrix data set the module associated with native cells, AFF2 (948 genes), was represented by *Srgn* (serglycin) and *Vamp1* (vesicle-associated membrane protein 1). In both accounts the biological process annotation 'response to stress' was significantly enriched.

Native tenocyte-associated modules

The modules with the strongest associations for native tenocytes, ILL4 and AFF3, demonstrated divergence in their hubs and annotation. ILL4 was found to contain troponins *Tnnt3* and *Tnni1*, *Kera* (keratocan) and *Gap43* (growth associated protein 43) as the most highly connected. The AFF3 module was represented by *Angpt2* (angiopoetin 2), *Robo4* (roundabout homolog 4), and the EMILIN-family member *Mmrn2* (multimerin 2). The module hubs differed in their annotations with the Illumina data defined by muscle contraction and muscle tissue development, whereas the Affymetrix module was described by blood vessel development and cardiovascular system development.

Monolayer-associated modules

Sample origins for monolayer differed in passage number with passage 5 for Illumina data and passage 3 for Affymetrix. The Illumina monolayer-associated module ILL2 had as hubs the beta-catanin inhibitor *Lzts2,* and regulators of growth plate differentiation *Igfr1* and *Sirt6*. The AFF4 module was found to contain *Fzd2* and *Bmp1* as central regulators. The Affymetrix module was

annotated with the terms 'regulation of cell-substrate adhesion' and 'extra-cellular matrix organisation', whilst the Illumina module was described by 'polysaccharide catabolic process'.

Alginate-associated modules

The Illumina alginate-associated module ILL1 contained the serine peptidase inhibitors *Serpina1* and *Serpina3n*, peptidase inhibitor *Pi15* and the transcription factor *Atf3*. The inhibitor of nuclear factor kappa-B kinase *Ikbke* was a hub in both the ILL1 and AFF5 alginate culture-associated modules. Additionally the AFF5 module contained *Ier2* (immediate early response 2) and *Sfrp2* (secreted frizzled-related protein 2). The ILL1 module was described by the term 'inflammatory response', whilst AFF5 was significantly enriched for the term 'positive regulation of cell differentiation'.

In summary, there was evidence for common functional annotations between modules from the two data sets using standalone analysis of these expression profiles, however, the qualitative preservation of hub gene candidates across phenotypic traits was equivocal nor consistently meet assumptions for confidently calling hub genes in some conditions.

Phenotypic Trait	Module association [cor, p-value]	Gene significance vs. Module Membership [cor, p-value]	Top 20 *k*ME – Hub genes	Biological Process (Top 50) annotation (adj.p-value)
Cartilage	**ILL6**/Turquoise [0.99, *p*=6e-30]	0.99, *p*<1e-200	Prg2, Nkg7, Add2, Napsa, Ppbp, Dnase1l3, Ctse, Plac8, Cfp, Fcnb, Loc24906, Ptprcap, Ifitm6, Ctsg, Camp, Sell, Retnlg, Ngp, Dmp1, Ms4a2	Immune system process (1.08e-7) Response to stress (1.3e-4)
Tendon	**ILL4**/Green [0.99, *p*=2e-29]	0.98, *p*<1e-200	Mb, Cox8b, Dhrs7c, Tnnt3, Myl1, Ryr1, Ckm, Myoz1, Tmod4, Itih3, Kera, Ccdc3, Art3, Tnni2, Rbfox1, Gap43, Eno3, Pgam2, Lyve1, Myoc	Muscle contraction (2.4e-16) Muscle tissue development (6.9e-6)
Monolayer	**ILL2**/Blue [0.91, *p*=2e-14]	0.83, *p*<1e-200	Scamp4, Srsf9, Apba3, Mrrf, Wsb2, RGD1559909, Dolk, Pcgf3, Leprel2, Prelid1, Igf1r, Ipo4, **Lzts2**, Clip2, Pygb, Prmt2, Slc30a5, Tm2d2, Sirt6, Eef1g	Polysaccharide catabolic process (9.6e-4)
Alginate	**ILL1**/Black [0.97, *p*=2e-21]	0.9, *p*=1.4e-30	Ptgds, Serpina1, Abcc9, Cesl1, Atf3, Gem, Ces1c, Gpr88, Pcsk1, Serpina3n, Ikbke, Map3k8, Rilp, Adipoq, Sectm1b, Maob, Ces1d, Akr1cl, Ppp1r1b, Pi15	Response to organonitrogen compound (3.6e-7) Inflammatory response (1.1e-4)
Models	**ILL5**/Red [0.95, *p*=2e-19]	0.88, *p*=1.1e-50	Armcx3, Ak3, Sat2, Eci2, P2rx4, Maoa, Pgrmc1, Commd10, Atp11a, LOC290595, Acsl4, Plin2, Psma1, Errfi1, Dnajc14, Sod2, Ctsd, Rbbp7, Tomm20, Mcfd2	Protein targeting to mitochondria (2.2e-4) Oxygen homeostasis (2.2e-4)

Table 4.3: Summary table highlights the phenotypic trait association with modules from **Illumina** data. Correlation of gene significance and module membership indicate confidence in modular hubs associated with traits. The top 20 genes with highest module membership (*k*ME>.9) as shown; the biological process annotation for these genes is also provided where significant annotations are present (GS>.5). The combined phenotype of 'Models' (both alginate and fibrin cultures) had a more robust association than fibrin alone and is presented. The gene *Lzts2* is highlighted as a monolayer hub – this is investigated further in **4.3.7**.

Phenotypic Trait	Module association [cor, p-value]	Gene significance vs. Module Membership	Top 20 *k*ME – Hub genes	Biological Process (Top 50) annotation (adj.p-value)
Native chondrocytes	**AFF2**/Brown [0.8, *p*=3e-6]	0.99, *p*<1e-200	Dnajb1, Mir29b2, F13a1, Hsph1, Pdk4, Pf4 ,Arid1b, Hsp90ab1, Isg20, Clec4d, Vamp1, Srgn, LOC363060, Fcer1g, Depdc7, Tcp11, Igsf6, Hspbap1, Cd53, Uspl1	Chaperone mediated protein folding (8.2e-6) Response to stress (8.7e-4)
Native tenocytes	**AFF3**/Green [0.95, *p*=3e-12]	0.59, *p*=9.1e-20	Gpr116, Cd93, Cdh5, Emcn, Tie1, Cd34, Myct1, Podxl, Eltd1, Prex2, Mfng, Angpt2, Npr1, Kit, Gpr4, Prkch, Tek, Robo4, Mmrn2, Plxnd1	Blood vessel development (4.6e-8) Cardiovascular system development (1.6e-6)
Native cells	**AFF2**/Brown [0.96, *p*=3e-13]	0.82, *p*<1e-200	Hubs based on module membership as per native chondrocytes above – gene significance value is higher for each gene for 'native' phenotype	NA
Monolayer	**AFF4**/Blue [0.69, *p*=2e-4]	0.5, *p*=8.4e-63	Loxl1, Adamtsl4, Flnc, Scrn1, Col8a1, Fzd2, Large, Lgals1, St5, Nkain1, Nuak1, Pcolce, Fam198b, St3gal2, Lhfp, Bmp1, Nxn, Ebpl, Ccbe1, Fbn1	Regulation of cell-substrate adhesion (8.4e-5) Extracellular matrix organisation (3.6e-4)
Alginate	**AFF5**/Red [0.92, *p*=2e-10]	0.42, *p*=3.6e-6	Ptprz1, Atp11a, Robo1, Etv1, Sfrp2, Tmem200a, Cyp7b1, RGD1310819, Nceh1, Ikbke, Itgb8, Fam168a, Tnfsf15, Acvr1b, Entpd4, Ptn, Ier2, Ppargc1a, Acadsb, Plekha5	JAK-STAT cascade involved in growth hormone signaling (6.1e-4) Positive regulation of cell differentiation (8.4e-4)
Model	**AFF6**/Yellow [0.8, *p*=3e-6]	0.3, *p*=1.8e-5	Galnt2, Junb, Efna4, Traf3ip2, Prickle2, Nudt4, Bst1, Naprt1, Lrp6, Pecr, Gnb5, Flrt2, P2rx4, Gpr153, Zmynd8, Shc1, Stxbp4, RGD1309534, Slc41a2, Sox4	Neurogenesis (8.7e-4) Cellular response to insulin stimulus (8.7e-4)

Table 4.4: Summary table highlights the phenotypic trait association with modules for **Affymetrix** data. Correlation of gene significance and module membership indicate confidence in modular hubs associated with traits. The top 20 genes with highest module membership (>0.9) are shown; genes are ordered (L-R) from highest module membership. The biological process annotation for these genes is also provided where significant annotations are present and are based on the top 50 genes. The relationship between the trait 'Model' and the yellow module was greater than that with 'Fibrin' and so the former is presented.

4.3.4: Consensus network generation for 3D culture systems

It is not possible to directly compare the networks features derived from individual data analysis, other that by functional annotation of the modules, as described above. In order to define the emergent properties in the three-dimensional culture systems it was necessary to undertake a differential eigengene network analysis to determine, and quantify, whether a set of modules found in one network was preserved in another. By pursuing this method a comprehensive analysis of network structures across two data sets would allow conserved and divergent functional modules to be defined to reveal biologically relevant pathway dependencies for a phenotype.

Expression data and eigengene network structure

To establish if the Illumina and Affymetrix studies were comparable the average expression rank for each gene in a data set was calculated and correlated between the two studies. The high expression correlation (cor = 0.7, p < 1e-200), **Figure 4.11,** suggested that the gene expression profile was significantly preserved between data sets. For module identification genes common to both data sets, after non-specific filtering, were identified (n=2795) and only these were used in further analysis. Soft-thresholding powers were tested across both data sets and $\beta = 7$ was chosen.

Consensus modules are identified from both data sets

Consensus modules are modules that are common to two gene co-expression networks and may represent biological mechanisms that are shared between the two data sources. Comparing the relationships between the consensus eigengene networks can reveal important differences between the systems. A consensus network was established by hierarchical clustering of a consensus dissimilarity

matrix. The consensus network analysis defined 15 modules, including one module ('grey' for unassigned genes). Additionally, new data set-specific single networks were prepared that were derived only from the 2795 common genes used for consensus network analysis, **Figures 4.12** and **4.13** (upper panels). As such, module identifiers used in sections **4.3.1-4.3.2** are not applicable to this section.

Module overlap with consensus modules is variable between data sets

Revised Illlumina and Affymetrix module eigengenes (derived from 2795 common genes) were considered against the 15 consensus modules to assess the gene overlap, i.e. whether genes comprising modules in the individual network analyses were wholly replicated in the consensus network. In general, there was overlap between module eigengenes from the individual networks and the consensus network, however, these overlaps were often not associated with a single consensus module eigengene, **Figures 4.12** and **4.13** (lower panels) and **Table 4.5**. Module overlaps are summarized in **Figure 4.14**.

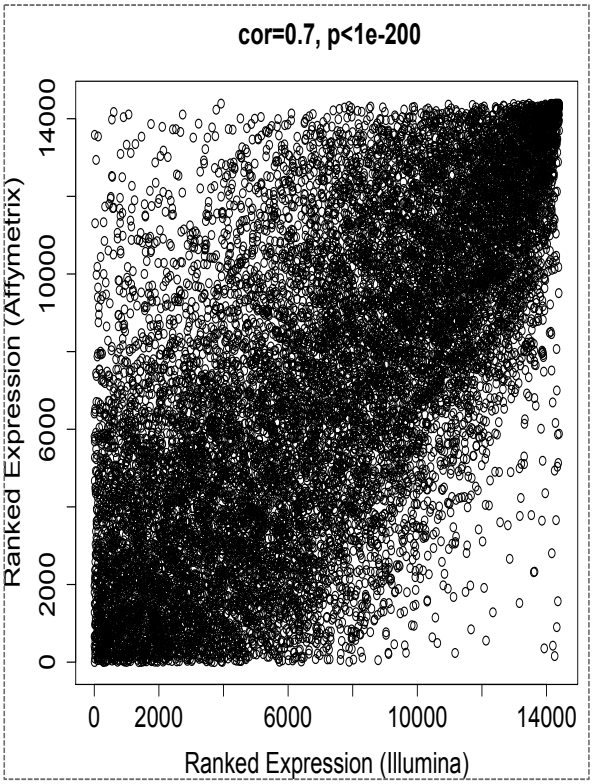
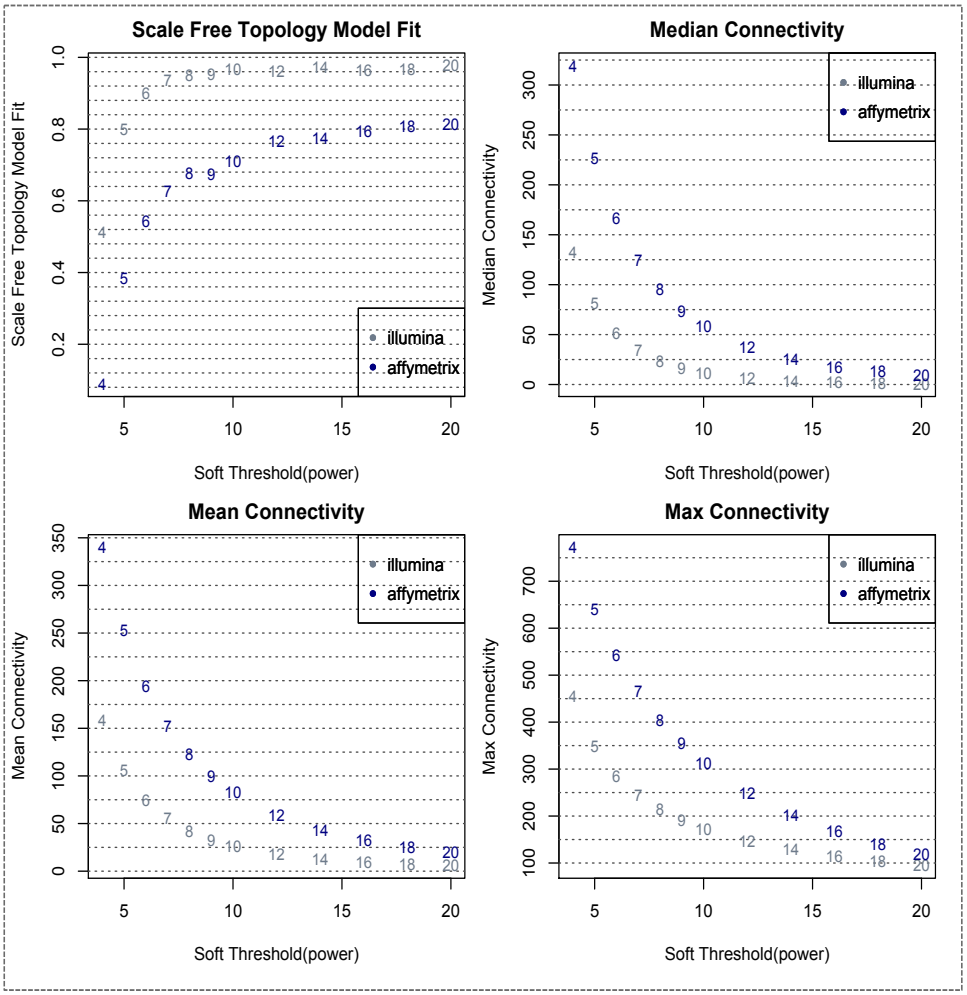


Figure 4.11: Upper panel - scatterplot of mean ranked expression values for every gene common to both Illumina (x-axis) and Affymetrix (y-axis) data sets (n=2975).

Lower panel: Soft-thresholding powers (x-axis) are plotted against scale-free topology criterion or connectivity values for either Illumina or Affymetrix data (boxed legends), y-axis.

Using the reduced gene numbers the scale-free topology issues arising in the earlier Affymetrix analysis are still evident. A power of beta=7 was chosen based upon the reduction in connectivity associated with increasing power in the Illumina data.



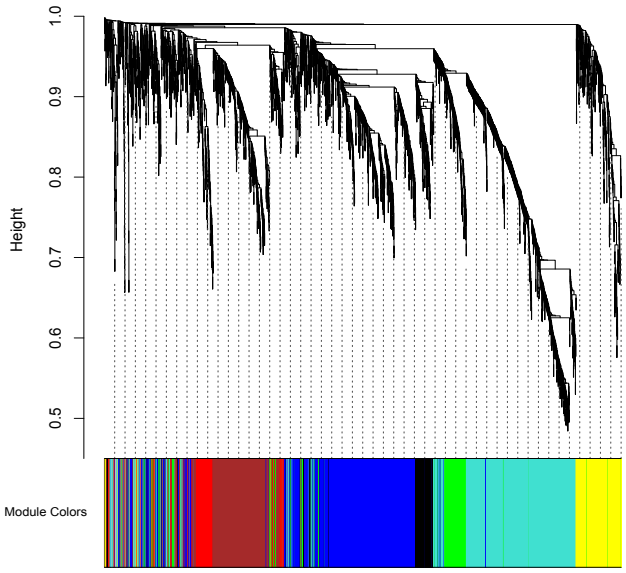


Figure 4.12: Upper panel - Clustering dendrogram for **Illumina** data using genes common to both data sets (top) with associated modules. General network structure is approximately comparable to that of the larger gene set. Nine modules were identified using the smaller data set (ILL8-ILL16) including the grey module, which contains unassigned genes. 2795 genes with deepSplit=1, beta=7, cut-height=0.2.

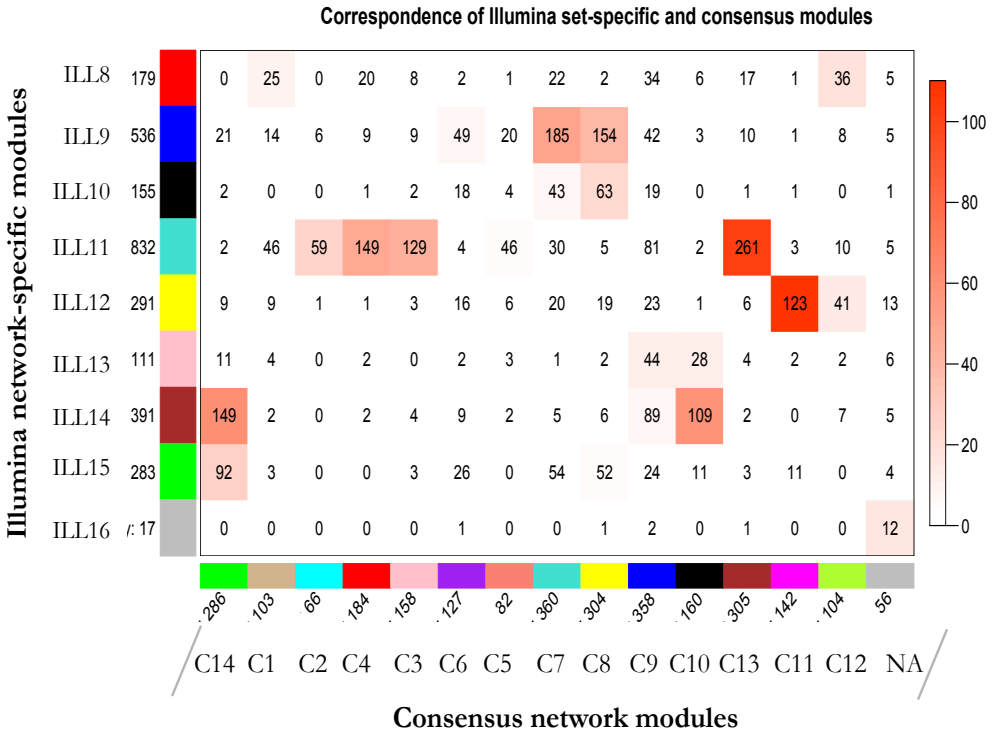


Figure 4.12: Lower panel - Gene overlaps between the revised Illumina modules (rows) and those defined by the consensus network for Illumina and Affymetrix (columns). Broadly, all Illumina network modules have representation in the consensus network, however, this may be over several consensus modules. Coloured cells define the number of genes from the consensus module that are overlapping with the Illumina-specific modules (linear heatmap). The number of genes in each module are indicated beside the colour blocks allocated to each module.

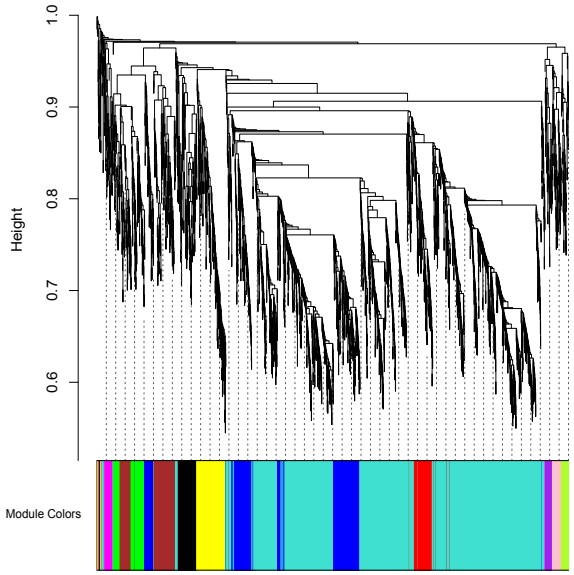


Figure 4.13: Upper panel - Clustering dendrogram for **Affymetrix** data using genes common to both data sets (top) with associated modules. General network structure is approximately comparable to that of the larger gene set. Eleven modules were identified (AFF7-AFF17) using the smaller data set. 2795 genes with deepSplit=1, beta=7, cut-height=0.2.

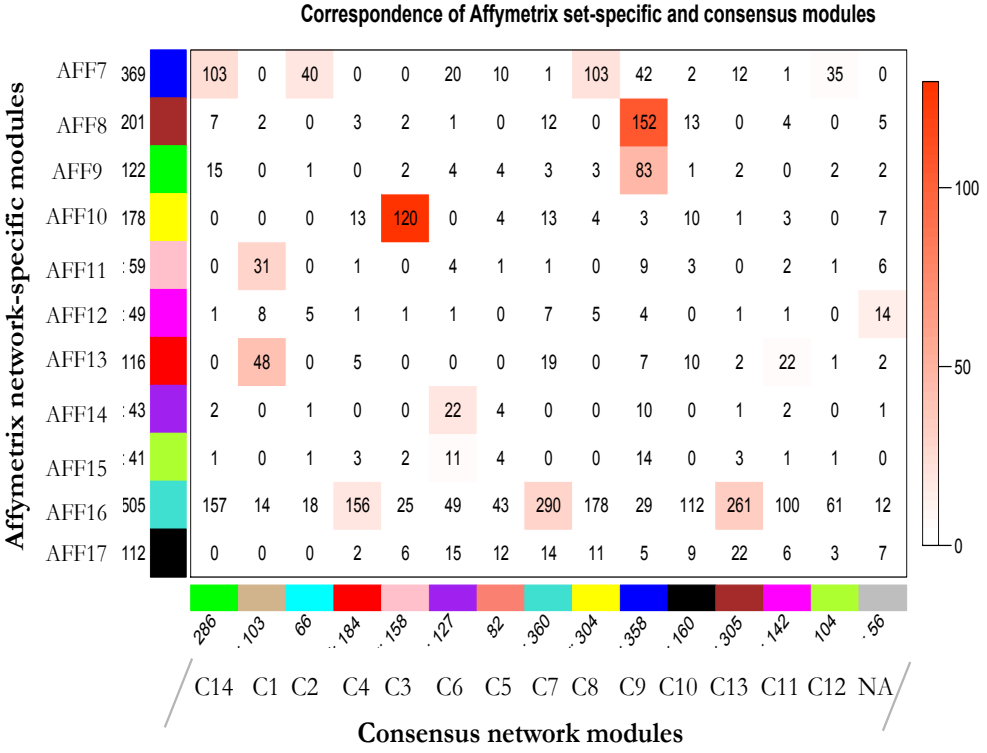


Figure 4.13: Lower panel - Gene overlaps between the revised Affymetrix modules (rows) and those defined by the consensus network for Illumina and Affymetrix (columns). Broadly, all Affymetrix network modules (AFF7-17) have representation in the consensus network, however, this may be over several consensus modules. Coloured cells define the number of genes from the consensus module that are overlapping with the Affymetrix-specific modules (linear heatmap). The number of genes in each module are indicated beside the colour blocks allocated to each module.

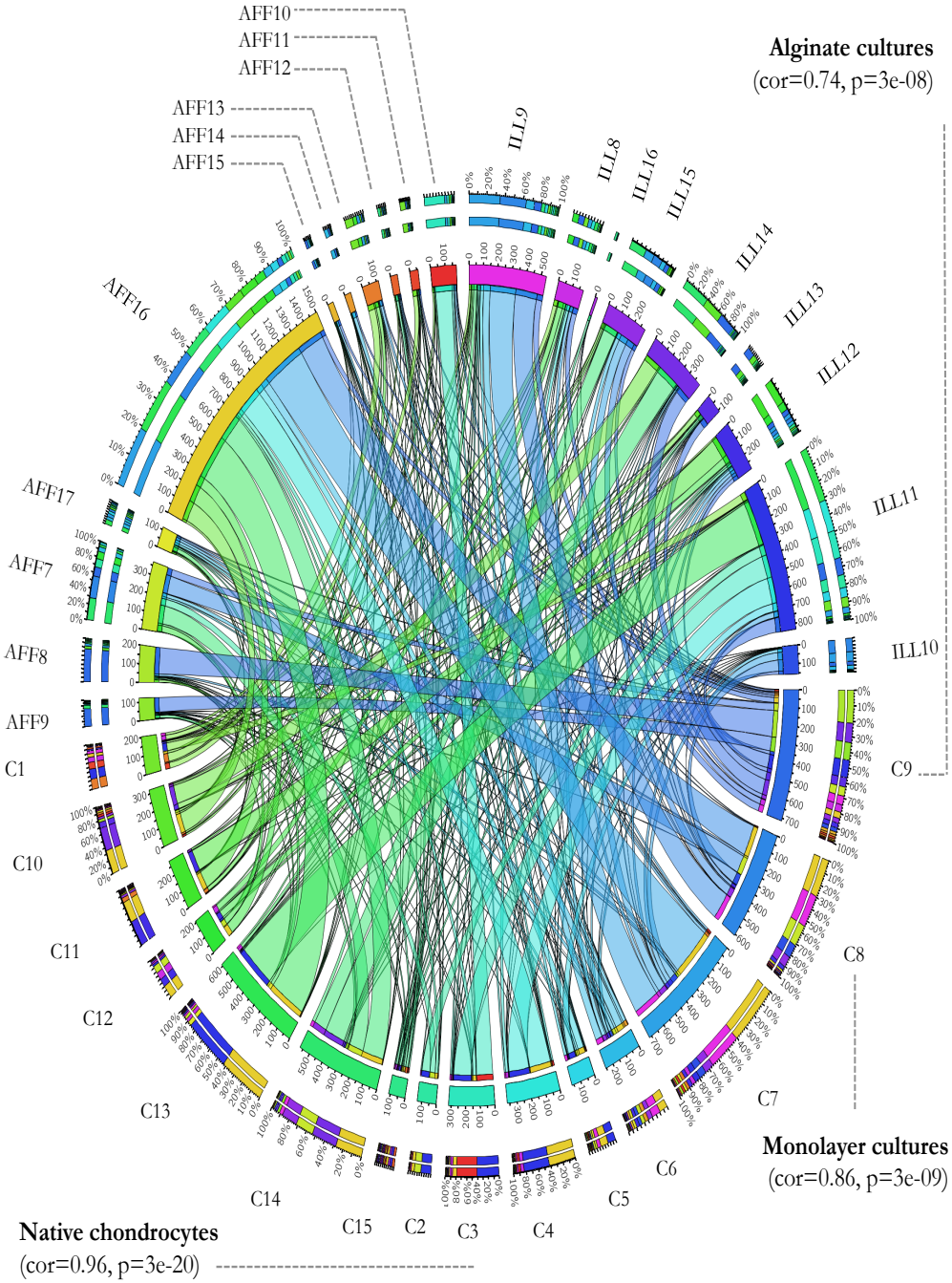


Figure 4.14: Circos plot to present Affymetrix ('**AFF**' prefix) and Illumina ('**ILL**' prefix) module overlaps with consensus modules (**C** prefix). Figure derived from tabular data where rows (Affymetrix or Illumina modules) and columns (consensus modules) are represented by coloured segments (inner circle) the size of which defines the total number of genes that overlap with the Affymetrix and Illumina modules. Ribbons connect rows and columns and are coloured by consensus module to show the overlap with each Affymetrix or Illumina module. The outer two rings define the relative contribution of each cell in a table to the row and column totals (stacked bar plots). The figure summarises the tabular data shown in **Figures 4.12, 4.13** and **Table 4.5**. N.B. The colours rendered have no association with the assigned module colours. The consensus modules with the highest trait associations for native and monolayer chondrocytes and alginate cultures are defined. The most highly preserved Illumina module in **Table 4.7** was ILL9, which together with AFF7, contribute the greatest number of overlapping genes with the C8, monolayer-associated, module.

Consensus Module (size)	Illumina Modules (overlap/illumina module size)	Affymetrix Modules (overlap/affymetrix module size)
C1: Tan (103)	**ILL8**/Red (25/179)	**AFF13**/Red (48/116)
C2: Cyan (66)	**ILL11**/Turquoise (58/832)	**AFF7**/Blue (40/369)
C3: Pink (158)	Turquoise (129/832)	**AFF10**/Yellow (120/178)
C4: Red (184)	Turquoise (149/832)	**AFF16**/Turquoise (156/1505)
C5: Salmon (82)	Turquoise (46/832)	Turquoise (43/1505)
C6: Purple (127)	**ILL9**/Blue (49/536)	**AFF14**/Purple (22/43)
C7: Turquoise (360)	Blue (185/536)	Turquoise (290/1505)
C8: Yellow (304)	Blue (154/536)	Blue (103/369)
C9: Blue (358)	**ILL14**/Brown (89/391)	**AFF8**/Brown (152/201)
C10: Black (160)	Brown (109/391)	Turquoise (112/1505)
C11: Magenta (142)	**ILL12**/Yellow (123/291)	Turquoise (100/1505)
C12: Greenyellow (104)	Yellow (41/291)	Turquoise (61/1505)
C13: Brown (305)	Turquoise (261/8320	Turquoise (261/1505)
C14: Green (286)	**ILL15**/Green (149/283)	Turquoise (103/369)

Table 4.5: Module overlap – A summary of the module overlap between **Illumina** or **Affymetrix** single network modules and the consensus modules. Consensus modules (**C**) are grouped into the meta-modules defined in **Figure 4.15**, with the exclusion of C13 and C14. For consensus modules the total number of genes in given in parentheses. For Illumina and Affymetrix modules the number of overlapping genes with consensus modules, relative to the total number of genes in the module, is provided in parentheses.

Only the greatest module overlaps are shown, complete overlap values are found in **Figures 4.12 and 4.13**. For data set-specific modules the reader is reminded that the allocation of colours to modules is arbitrary and so, for example, the 'brown' Illumina module is not necessarily equivalent to the 'brown' Affymetrix module.

Module equivalence is defined by the consensus module overlaps and statistical definitions in **Table 4.7**. To aid discussion alternative alphanumeric names are given to the consensus modules ($C1 \ldots C_n$).

4.3.5: Differential eigengene network analysis and meta-modules

The analysis proceeded by summarizing consensus modules for each data set by the first principal component, the module eigengene. As eigenegenes from different modules show correlations eigengene networks can be defined. This may show whether the identified consensus modules that are highly related in one data set (Illumina) are also highly related in another data set (Affymetrix).

An eigengene network was constructed for each data set such that highly correlated eigengenes for different modules were grouped together, **Figure 4.15:i-ii**. For each module eigenegene the scaled connectivity (degree) was defined as the mean connection strength with other eigengenes (**Figure 4.15:iv**). The average scaled connectivity across the whole eigengene network was defined as the density, *D*. If most eigengenes within a network have high, positive correlations with each other the value of *D* approaches 1, i.e. the eigengene network relationships are highly comparable between the data sets.

Between the eigengene networks of consensus modules for Illumina and Affymetrix data sets there was moderate preservation as defined by the density value, $D = 0.68$. The relationship of each individual eigengene with all others in the network was variable, reflected by the differences in individual module scaled connectivity values in the preservation network (**Figure 14.5:iv**.). The C9 (blue) module eigengene was found to have the highest scaled connectivity in the preservation network.

As eigengenes form networks analogous to the earlier analysis modules comprised of eigengenes, 'meta-modules' can be identified, **Figure 14.5: iii and vi**, which show high, positive correlations with each other. Meta-modules can demonstrate

higher order relationships in the gene co-expression organization that are not evident in standard module detection. In the Illumina data the module eigengene network was clustered in four meta-modules: meta-module C1:C4 - tan, cyan, pink and red; C5:C8 - purple, salmon, turquoise and yellow; C9:C10 - black and blue; and C11:C12 - greenyellow and magenta. Whilst there was some evidence of the conservation of meta-modules in the Affymetrix data, especially C1:C4 and C5:C8, preservation values were often poor. The meta-modules were functionally annotated and are summarized in **Table 4.6**.

Module preservation analysis

Module overlap, and the definition of module equivalence between data sets, was not clear. To define the preservation of Illumina modules (considered the reference data set) in Affymetrix modules (test data set) statistically a permutation test was employed. Module preservation was generally low, however, the ILL9, ILL15, ILL13 and ILL10 Illumina modules were the most significantly preserved in the Affymetrix network, **Table 4.7**.

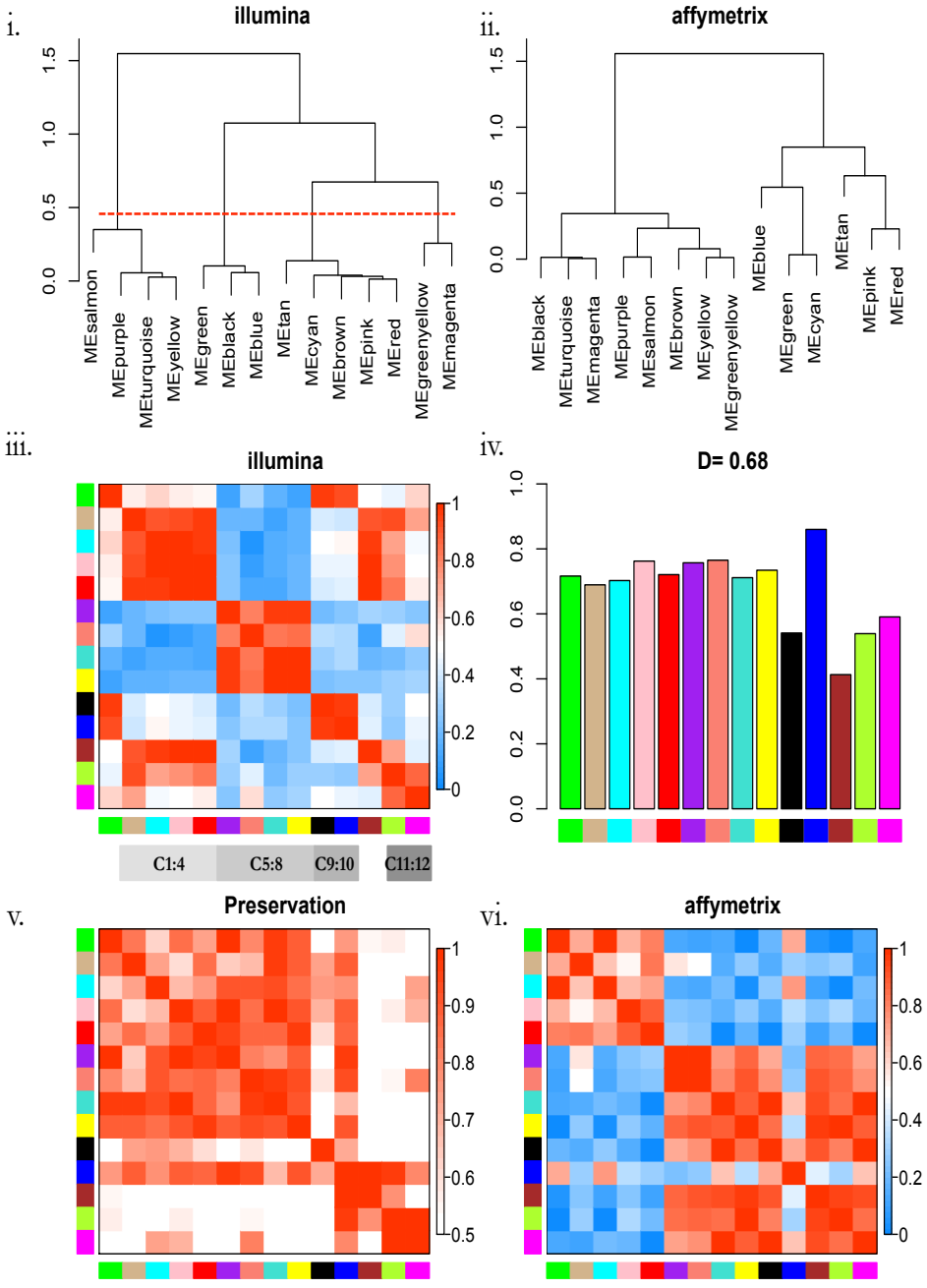


Figure 4.15: Differential eigengene analysis in Illumina and Affymetrix data sets - **i, ii**. - dendrograms of consensus module eigengenes for Illumina and Affymetrix networks; **iii**. - heatmap of eigengene adjacencies for the consensus eigengene network relative to the Illumina data. Four meta-modules are evident in **i**. and **iii**. Rows and columns relate to one consensus eigengene (labelled by the consensus module colours) – positive correlation/high adjacency is shown by red intensity; blue represents negative correlation as defined by the vertical colour bar; **vi.** – represents the equivalent plot for the Affymetrix data; **v**. – histogram of the preservation values for each consensus eigengene (bar colour) – height (y-axis) represents the eigengene preservation measure. *D*, density, represents the overall eigenegene network preservation; **v.** – heatmap of preservation network adjacencies with figure structure as per **iii**. Generally, the consensus module eigenegene network preservation is moderate between the data sets with limited replication of meta-modules in Affymetrix data.

Table 4.6: Consensus modules are gathered into meta-modules (**Figure 4.15**) and the overall trait association is presented. Using all the genes within the meta-modules the gene ontology analysis is shown for biological process, metabolic function and cellular compartment using DAVID with a general *Rattus norvegicus* background. Modules below the black line show no consensus associations with phenotypic traits.

Consensus Meta-module	Phenotypic Trait	Biological Process	Metabolic Function	Cellular Compartment	KEGG pathway
C1-C4 Tan Cyan Red Pink	Native cells	Transmembrane receptor protein serine/threonine kinase signaling pathway (1.9e-3) Anatomical structure formation involved in morphogenesis (2.4e-4) Developmental process (1.8e-6) Extracellular matrix organisation (9.8e-3)	Actin binding (4.8e-3)	Extracellular matrix (1.9e-3) Vesicle (4.9e-3)	TGF-beta signalling pathway (9.8e-3)
C5-C8 Purple Salmon Turquoise Yellow	Monolayer	Response to stress (1.6e-8) Defense response (6.7e-5) Biological adhesion (1.8e-3) Response to organic substance (1.6e-4)	Catalytic activity (1.7e-9) Oxidoreductase activity (1.9e-5)	Cytoplasm (4.9e-4) Extracellular region (1.3e-2)	NA
C9-C10 Black Blue	Model cultures	Developmental process (2.5e-7) Cell differentiation (6.1e-5) Response to stress (1.9e-8) Response to inorganic substance (3.2e-7)	Catalytic activity (5e-6)	Extracellular region part (4.3e-12) Extracellular matrix (8e-6) Lytic vacuole (3.9e-3)	Lysosome (3.9e-2)
C11-C12 Magenta Greenyellow	No consensus association	Cholesterol biosynthetic process (1.1e-3) Cellular process (1.9e-3)	Ion binding (4.1e-3) Insulin-like growth factor binding (2.1e-2)	Cytoplasm (4e-13) Proteinaceous extracellular matrix (5.2e-3)	Steroid biosynthesis (4.5e-2)
C13: Brown	No consensus association	Cell cycle phase (8.9e-10) Microtubule-based process (2.6e-5)	Microtubule motor activity (ns)	Intracellular organelle part (3.1e-12)	DNA replication (4.7e-9)
C14: Green	No consensus association	Cellular metabolic process (2e-3) Response to organic substance (1.4e-2)	NA	Cytoplasm (2.4e-12) Mitochondrion (9.8e-5)	NA

Illumina Module	Module size	Z-score, Preservation summary	Log_{10} p-value, Bonferoni
ILL9/blue	400	13.29	-42.81
ILL15/green	283	7.51	-16.22
IL13/pink	111	7.24	-17.02
ILL10/black	155	5.17	-6.66
ILL11/turquoise	400	4.73	-5.94
ILL8/red	179	4.51	-6.89
ILL14/brown	391	4.49	-6.09
gold	100	2.82	-2.71
ILL12/yellow	291	-0.39	0
grey	17	-0.85	0

Table 4.7: Module preservation - To quantify how well the modules in the Illumina (reference) network were preserved in the Affymetrix (test) network a permutation test was employed. The maximum module size used for calculations was 400 genes; modules larger than this consisted of random samples of genes within the specified module.

In general terms, the higher the Z-score summary preservation value the more preserved the module is between data sets. Values 5>Z<10 represent moderate preservation, whilst Z>10 indicates high preservation. Uncharacterised genes are within the 'grey' module, whilst the 'gold' module is generated as part of the test and contains randomly assigned genes from any module; 'grey' and 'gold' modules should score consistently lower than preserved modules. The ILL9 module (**Figure 4.12**) shows the strongest preservation

Phenotypic Trait	Module association [cor, p-value]	*k*ME correlations	Top 20 *k*ME
Monolayer	ILL9/blue [0.88, *p*=1e-12]	0.59, *p*=1.4e-51	Rgd1566262, Prelid1, Chpf, Lmf2, Wsb2, Lass5, Geft, **Itga11**, **Gpc4**, Ccs, Pofut2, Pygb, Mrpl55, Lman1, C1qtnf5, Stub1, Rab30, Ipo4, Rgd1559909, Smyd2, Tagln
Alginate	ILL13/pink [0.97, *p*=3e-22]	0.64, *p*=4e-14	**Gpnmb**, Cp, Orm1, Sectm1b, **Chi3l1**, Obfc2a, **Serpina3n**, Crispld2, Enpp2, Atf3, Map3k8, Zfp347, Coro6, **Pi15**, Btg2, Cd302, Maoa, C3, Phlda1, Cfh

Table 4.8: Top consensus hubs based on high *k*ME values across networks – only the ILL9 and ILL13 modules shared significant *k*ME preservation (**Figure 4.17**) with the Affymetrix network. These modules were associated with the 'monolayer' and 'alginate' traits for Illumina modules. For monolayer the consensus hubs included glypican 4 and integrin alpha 11; for alginate peptidase inhibitors *Serpina3n* and *Pi15,* and glycoprotein *Gpnmb* were called as consensus hubs.

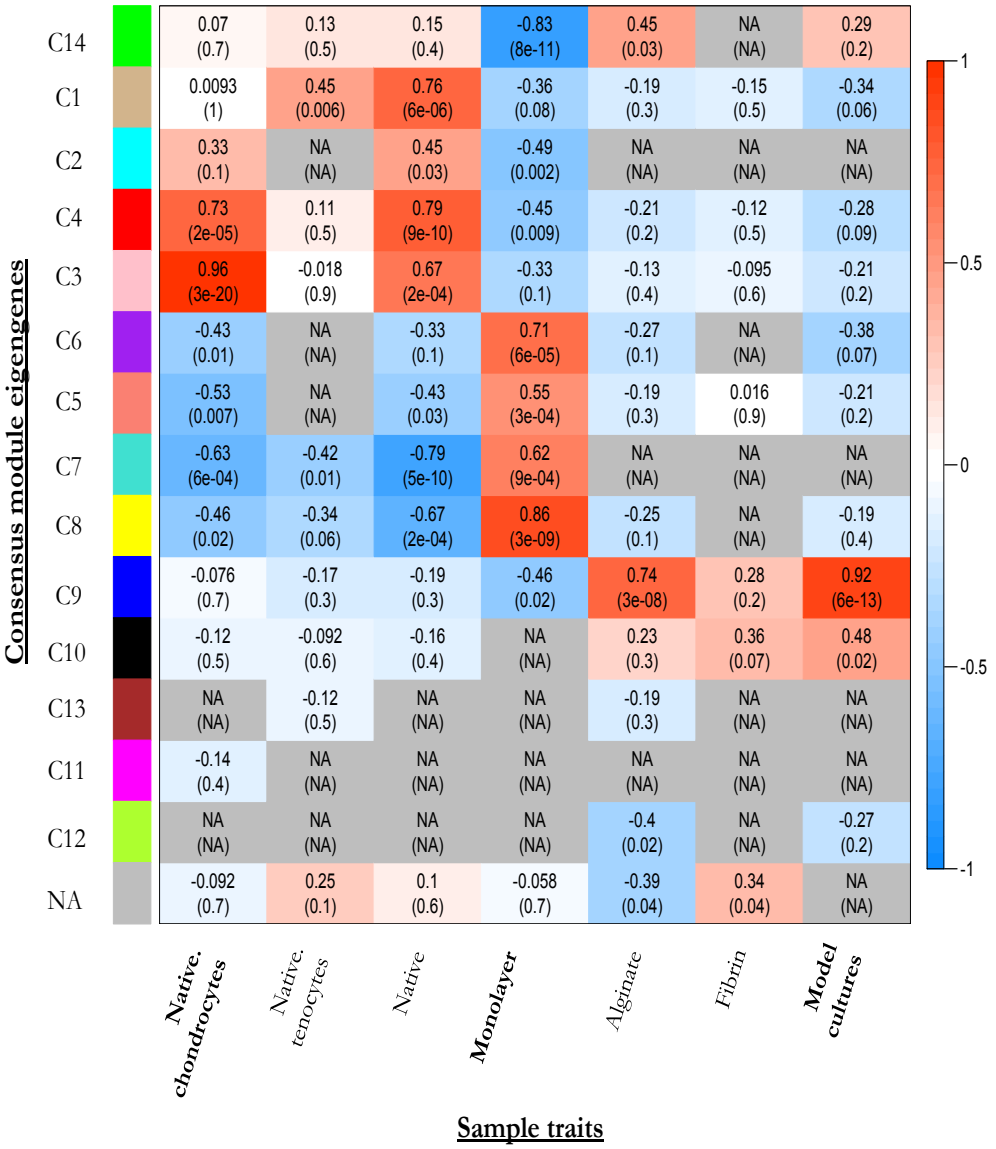


Figure 4.16: Summary relationship between the **consensus** module eigengenes (rows) and common conditional traits (columns) for Illumina and Affymetrix data sets. Within each matrix cell the correlations between the the corresponding module eigengene and the trait is reported. The *p*-values are provided in parentheses below. The cell colour represents the correlation value (red - positive, blue - negative) and is coded in the vertical colour bar on the right of the figure. Some traits combine terms, for example, 'Native' contains both native chondrocyte and tenocyte data, whilst 'Model cultures' contains both alginate and fibrin samples. Missing (NA) cells indicate that correlations in the module-trait pairs in the Illumina and Affymetrix data sets had opposite signs and no summary relationship can be defined. Native chondrocytes were generally associated with modules in the pink|red|cyan meta-module with the pink module showing the greatest association with the trait. Strong consensus relationships were not found for tenocytes, but their inclusion in the 'Native' phenotype improved association with the tan and cyan modules. The 'monolayer' trait was associated with the purple|salmon|turquoise|yellow meta-module with the yellow module having the strongest association with this trait. There was no strong association with any module for the trait 'fibrin', but when aggregated with the alginate samples the trait 'model cultures' was strongly associated with the blue consensus module eigengene.

Considering the module-trait associations with the functional annotations of the meta-modules validated the annotations found in the two data sets. Monolayer, for example, was associated with the C5:C8 meta-module and the functional terms 'developmental process' and 'anatomical structure formation involved in morphogenesis'; model cultures, alginate and fibrin, were defined by the C9 and C10 meta-module and the terms 'developmental process' and 'cell differentiation'. These functional annotations have been consistent in the analysis across both the Illumina and Affymetrix data sets.

To determine common hub genes across studies, which may act as central regulators, Module Memberships were calculated for all genes using the module eigengenes for the Illumina data. The Module Membership value, *k*ME, is useful in comparing networks as it is a global measure of the correlation between every gene and each module eigengene. This allows genes not initially found to be assigned to a module to be considered as central regulators between networks. In essence this method superimposes the modular structure (colour assignments) of one network onto another.

Module Membership values for every gene in the Illumina data set were considered against the equivalent gene *k*ME in the Affymetrix data. Only the Illumina ILL13/pink (0.64, p = 4e-14) and Illumina ILL9/blue (0.59, p = 1.4e-51) modules demonstrated high correlation between the *k*ME values indicated preserved module connectivity, **Figure 4.17**. On this basis only these two modules will be considered further.

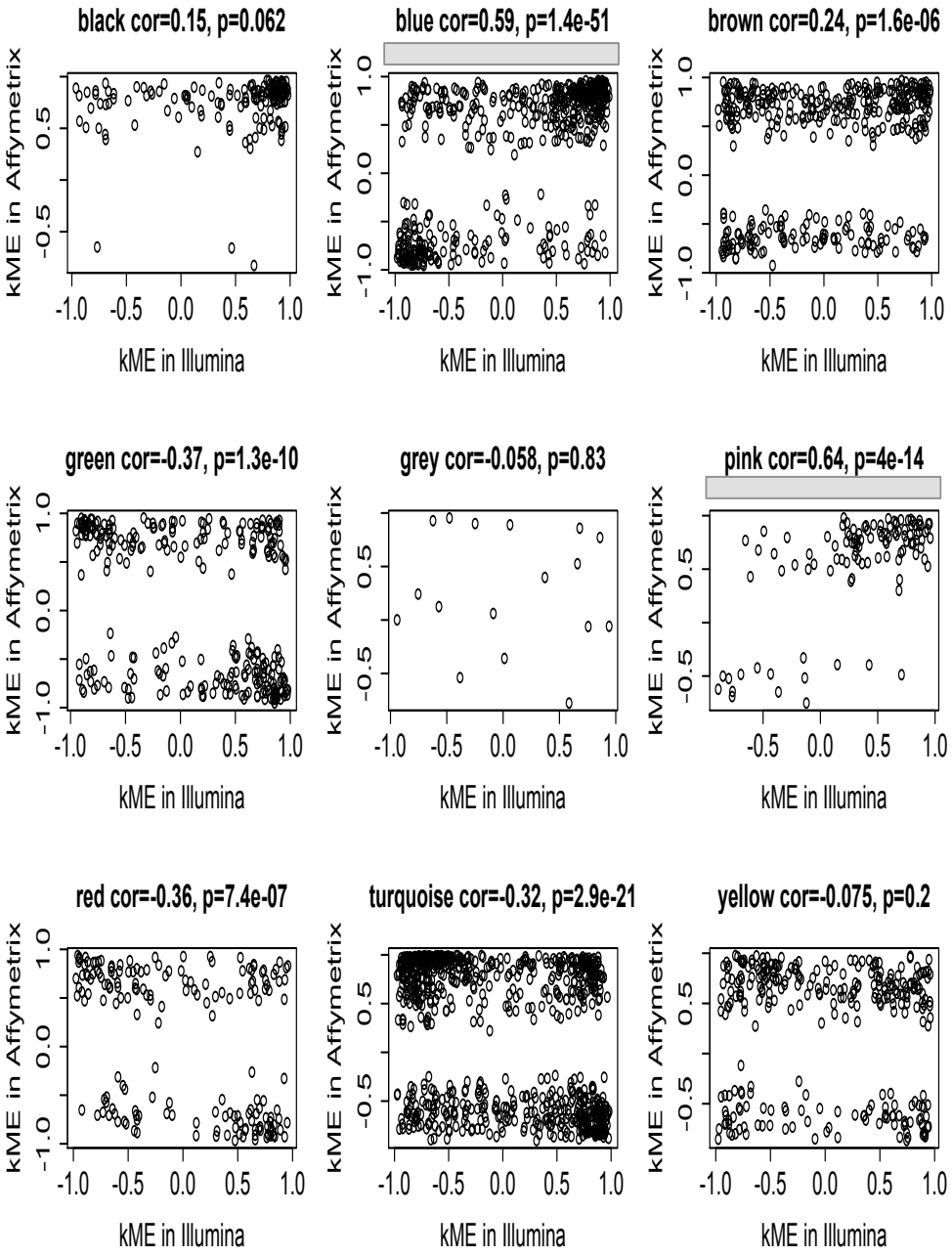


Figure 4.17: Scatterplots of *k*ME Illumina (x-axis) against *k*ME Affymetrix (y-axis). Using only in-module genes is a visual way of assessing hub gene conservation: if these genes show between-set correlation, then the genes in the upper right of the plot are likely to be common hub genes between data sets. Hub genes are genes that show significant correlation with module eigengenes and high within-module connectivity. Module membership values (*k*ME) were calculated for all genes in both data sets using the Illumina module classifiers as the reference network. Only conserved hubs should have high *k*ME value correlations when the test network (Affymetrix) is compared to the reference. Only the Illumina ILL13 (pink) and ILL9 (blue) modules (highlighted with grey boxes) were found to have this association. In other terms, confidence in calling consensus hub genes could only be extended to genes from these two modules.

To define which genes were hubs in both datasets gene lists were ranked on their module membership and the top twenty selected, **Table 4.8**. The common central regulators found with the ILL13 alginate-associated module, included *Atf3*, the peptidase inhibitor *Pi15*, chitinase *Chi3l1* and the regulator of osteoblast-osteoclast differentiation and function osteoactivin/*Gpnmb* and the serine protease inhibitor *Serpina3n*. The ILL9 module, monolayer-associated, contained the common hubs *Itgna11* (integrin α11), *Gpc4* (glypican 4) and *Tagln* (transgelin).

To qualify these findings for alginate cultures the modular network for ILL13/pink was visualized using Cytoscape with either the Illumina or Affymetrix gene expression data from **Chapter 2** and **3**. The topological overlap matrix of either data set was used as the basis for the network and the gene members of the ILL13 module were the nodes around which the network was built. In **Figure 4.18** the consensus network hubs that show the strongest intramodular connectivity are *Pi15* and *Serpina3n*; few of these consensus hubs show evidence of protein-protein interactions, **Figure 4.19**. In **Figure 4.20** the topological overlap of the Affymetrix data on the ILL13 module also placed *Serpina3n, Pi15*, along with *Chi3l1*, as modular hubs, which were more highly expressed in alginate cultures relative to either monolayer or native chondrocytes.

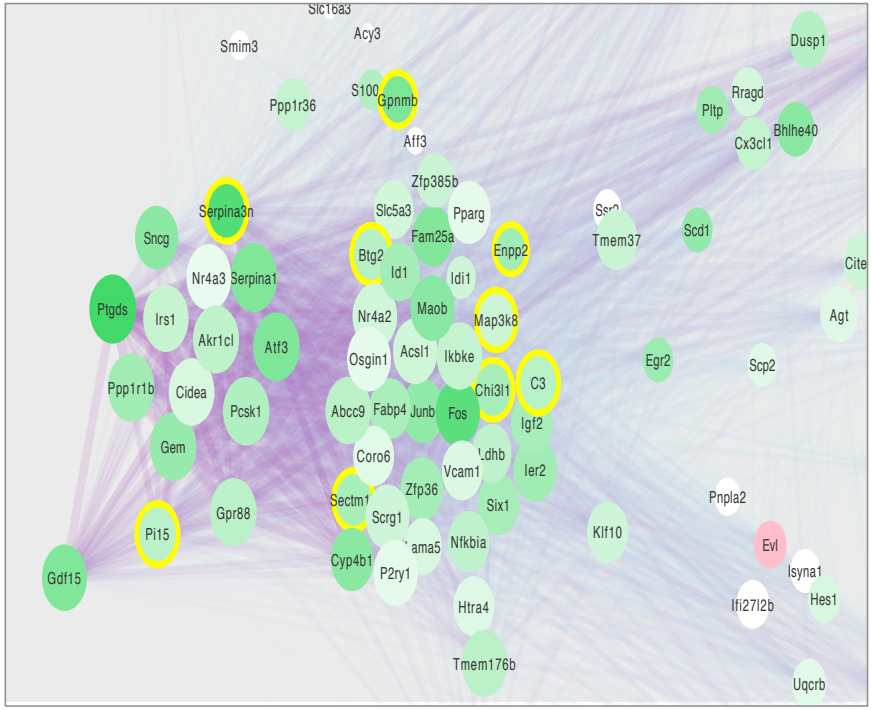


Figure 4.18 Graphical representation of topological overlap of the Illumina ILL13/pink module. Nodes represent genes within the module and **Chapter 2** $\log_2$ fold-changes for the native cartilage to alginate differential expression analysis are overlain such that genes more highly expressed in cartilage are red (**figure legend**). Genes identified as consensus hub genes are highlighted by a yellow corona. From the Illumina data the topological overlap would indicate that of the consensus hubs *PI15* and *Serpina3n* have the highest intramodular connectivity The network structure may be compared against the putative hub genes defined in the single network analysis for Illumina data in **Table 4.3**.

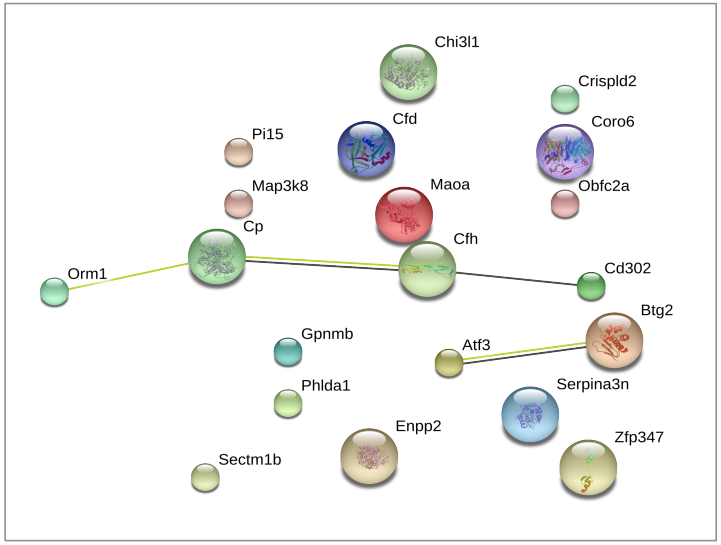
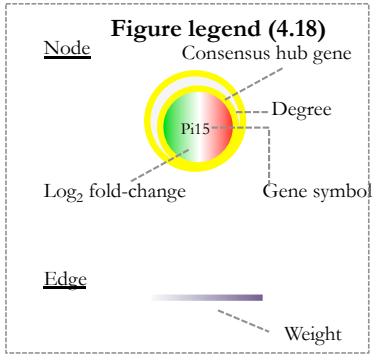
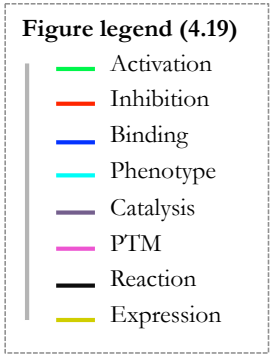


Figure 4.19: Output from STRING to show the protein-protein interaction (PPI) evidence view for the top 20 consensus hub genes identified as having strong association with chondrocytes in alginate beads. Using this database there is no evidence for PPI between Serpin3an, Pi15 and Gpnmb.





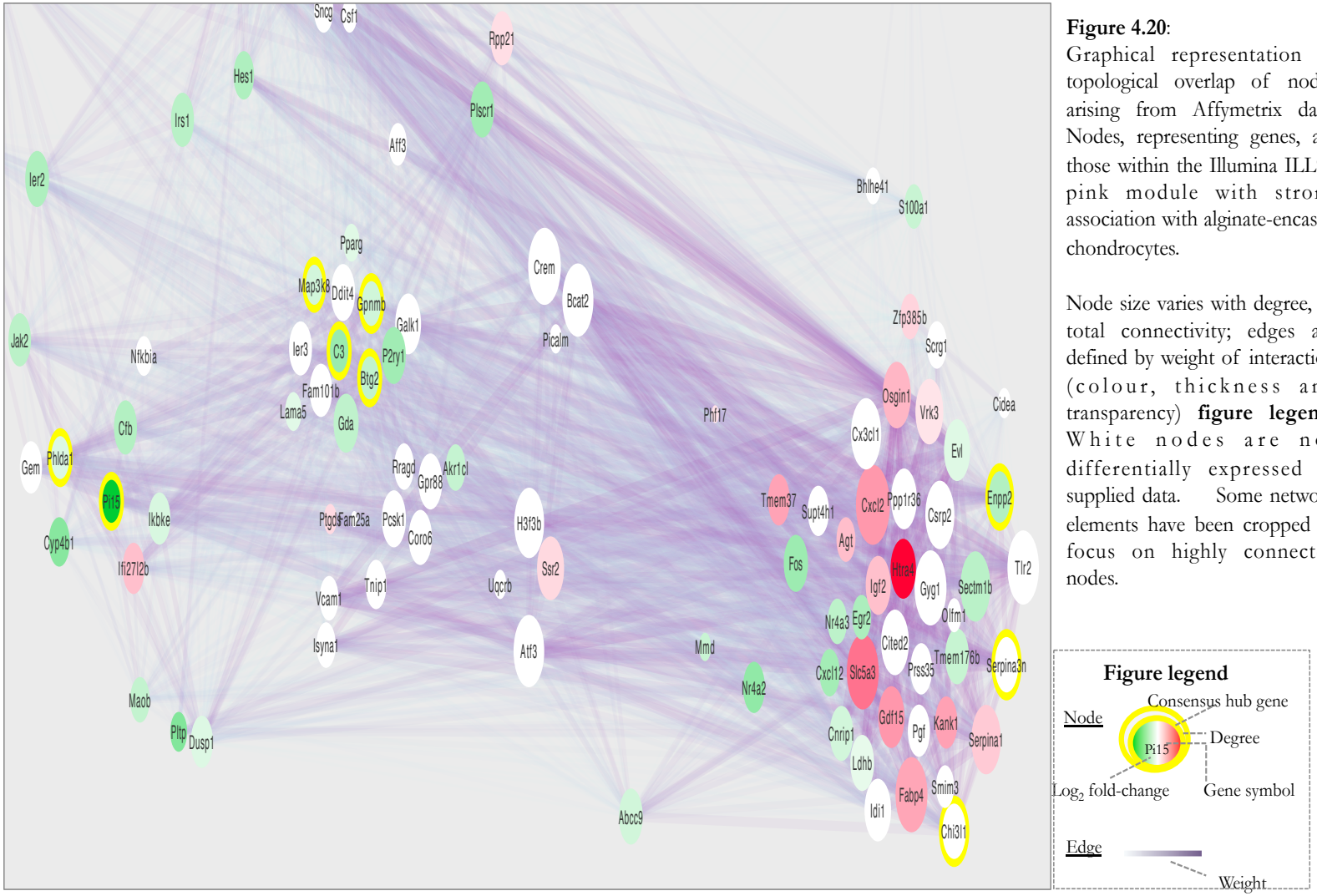


Figure 4.20:
Graphical representation of topological overlap of nodes arising from Affymetrix data. Nodes, representing genes, are those within the Illumina ILL9/ pink module with strong association with alginate-encased chondrocytes.

Node size varies with degree, or total connectivity; edges are defined by weight of interaction (colour, thickness and transparency) **figure legend**. White nodes are not differentially expressed in supplied data. Some network elements have been cropped to focus on highly connected nodes.

4.3.7: *Lzts2* silencing does not influence chondrocyte differentiation status

In the independent analysis of the Illumina data the leuzine-zipper, putative tumour suppressor 2, *Lzts2,* was identified as a potential hub or central regulator within a module showing strong correlation with the monolayer trait for chondrocytes and tenocytes. Expression was found to be lower in both monolayer chondrocytes (log_2 fold change = -1.56; false discovery rate = 1.14e-08, log odds ratio = 12.01) and in tenocytes at passage five (log_2 fold change = -0.86, false discovery rate = 1.4e-4, log odds ratio = 2.9) relative to native tissue (**Chapter 2**). To determine whether *Lzts2* was a central regulator in the dedifferentiation of chondrocytes a siRNA approach was employed to silence *Lzts2* expression and evaluate the impact on differentiation markers by qPCR. Based upon differential expression analysis *Col2a1* and *Acan* were selected as markers of differentiated status; genes more highly expressed in dedifferentiation *Pitx1* and *Thy1*, and the chondrogenesis regulator *Sox9*, were also considered.

Robust reduction (~77% knockdown) of expression of *Lzts2* was consistently shown in chondrocytes at passage three, **Figure 4.21**. Trending differences in expression between *Lzts2* siRNA treated and negative control samples, for aggrecan (*Acan*), were confounded, however, by higher expression of target genes in some samples only treated with the transfection agent, Lipofectamine 2000. Two independent studies could not demonstrate clear evidence of an impact of *Lzts2* knock-down on marker of chondrocyte differentiated status (data not shown).

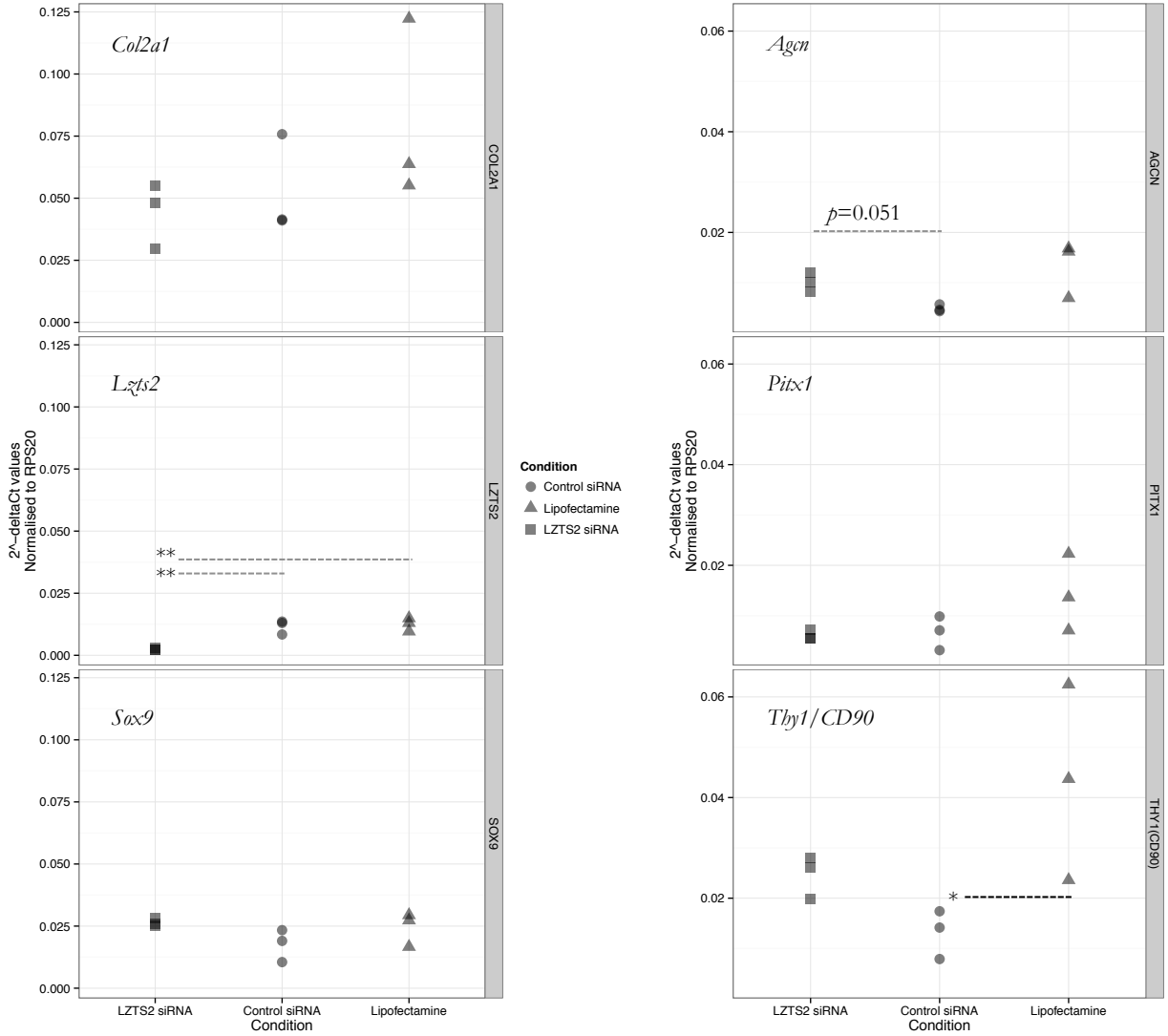


Figure 4.21 – Plots of quantitative PCR data (2^-dCt, y-axis) for three treatment conditions (x-axis) for passage 2 chondrocytes – i) treated with *Lzts2* siRNA, ii) treated with negative control siRNA, iii) treated with comparable volume of Lipofectamine 2000. Significance code: (*) – p<0.05; (**) – p<0.01

Expression of the leucine zipper *Lzts2* was significantly reduced in cultures following application of siRNA targeted to this transcript (p<0.01).

There was also a significant difference in expression between the negative control siRNA-treated chondrocytes and the lipofectamine controls for *Thy1*.

Differentiation (*Col2a1*), dedifferentiation (Pitx1, Thy1) and development markers (Sox9) were not significantly different between groups. Differential expression of aggrecan gene (*Agcn*) trended towards significance (p=0.051). All samples were normalised to *Rps20*.

4.4: Discussion

Weighted gene co-expression network analysis (WGCNA) is a systems biology methodology that facilitates investigation of the global network properties of a transcriptome and provides functional insights into the organization of the network. The practical output of this type of analysis is the elucidation of higher-order relationships between groups of highly co-expressed genes (modules) and their associations with phenotypic traits. Comparative analysis of network structures allows the definition of conserved and divergent functional modules between data sets providing a global network perspective of changes in the system under investigation.

Differential expression analysis in **Chapters 2** and **3** focused on the selection of individual genes as candidates for modulation of differentiation status in chondrocytes and tenocytes. The gene expression profiles defined in these earlier studies arose from the collective responses of multiple regulatory networks and this information is lost in single gene analysis (Zhao, Langfelder et al. 2010). In order to determine regulatory units that may modulate differentiation status in chondrocytes and tenocytes the WGCNA approach was employed to consider the connectivity of all genes analysed and derive co-expression networks that describe the strength of relationships between genes and their topological overlap with other genes in the network.

As outlined in **Chapter 3** the transition to an alternative microarray platform negates the direct comparisons between the gene expression profiles derived from Affymetrix and Illumina studies. Comparison of the global network properties using WGCNA overcomes this limitation and can facilitate the definition of

regulatory sub-networks driving phenotypic changes that was not otherwise possible using data arising from two different platforms. The hurdle of integrating gene expression profiles across platforms is explored further in **Chapter 5**.

4.4.1 Application of methodology

The practicalities of this type of methodology should be considered and whether it readily improves interpretation of gene expression data beyond what is possible by standard differential gene expression analysis alone.

To identify consistent and unambiguous behavior of putative hub genes is likely to be difficult. Langfelder, *et al* (2013) stated that in the majority of cases global hub genes are not important and the focus should be on trait-associated intra-modular (sub-network) hubs, those highly interconnected within a functional gene aggregate (module). Furthermore, gene selection strategies based on networks are not necessarily any more reproducible in terms of their trait associations that standard approaches such as differential gene expression when it comes to validation (Langfelder, Mischel et al. 2013). The question arises as to whether candidates selected from hub gene lists any more likely to be predictive biomarkers than those derived from standard differential expression or meta-analysis techniques. In a recent example Yang, *et al* (2014) found cancer genes with prognostic potential lay within modules, i.e., they were not global hubs (Yang, Han et al. 2014).

Co-expression analysis may be influenced by a number of extraneous systematic differences between data sets that may result in the discovery of erroneous correlations (Gaiteri, Ding et al. 2014). Therefore, feting intra-modular hubs from single data sets analysis may be dangerous. The definitive performance criterion

would be the evidence of preservation of trait associations and hub genes from modules in independent datasets. Langfelder, *et al* (2013), as before, showed that use of putative biomarkers derived from marginal meta-analysis often outperforms selection of intra-modular hub genes in prediction testing; however, in noisy, heterogenous data with weak signals, but strong module membership preservation (as is the case in the data sets presented here) the inverse is true. In practical terms, the wider significance of predicted regulatory hubs in isolated data sets still requires substantial validation, as discussed further below.

4.4.2: Genes identified as hubs in single network analysis are not reproduced in consensus network analysis

As part of the WGCNA analysis global gene expression profiles from two data sets were initially analysed independently. Hub genes in single data sets were selected on the basis of high module membership (*k*ME), which was strongly related to intra-modular connectivity. The decision to investigate the relevance of one hub gene (*Lztś2*) over others was defined by the findings in the literature, discussed below, and financial restrictions on investigating each identified hub.

Hub genes from single network analysis were not reproducible in the consensus analysis, i.e. trait-associated modules did not show overlap in the predicted hubs across Affymetrix and Illumina studies. Common hubs with high *k*ME were identified between the two networks as part of the consensus network analysis, but the veracity of these findings is questionable given the poor preservation of module eigengenes across the two networks.

It is not clear why the network structures, and therefore also the hubs, were so poorly preserved across the data sets. There were multiple modules associated

with each trait, which can complicate the selection of hubs. Further to this, robust definition of intra-modular hubs is critically related to their identification across very different datasets and, as such, the definition of module membership is only relevant and successful where the module is actually present across data sets (Gaiteri, Ding et al. 2014) and a reproducible relationship exists with a trait. Although there was high correlation of mean rank expression across these two datasets the fact that they arose from different platforms and differing experimental designs may have had a critical bearing on the ability to find consensus modules related to the various phenotypes. On this basis, it is not possible to confidently state that the hubs identified using this methodology on these data sets has identified more reproducible markers of de- and re-differentiation than those already highlighted in the standard differential analysis in **Chapter 2** and **3** without further validation.

Beyond the hub: are module hubs relevant?

There are multiple approaches to defining co-expression networks, functional modules and selecting hub genes. There appears to be no systematic analysis of these techniques, or indeed of publications that have used them, to determine how robust they are across datasets. In many cases, and this cannot be considered an exception, the complexity of the approaches relegates each technique to a 'black-box' - the emergent hubs may have little relevance beyond the context of the study. Whether similar hubs are found across divergent studies would be of interest.

Co-expression links between two or more genes are made without reference to the molecular mechanisms through which the mRNA (or protein) came to be expressed. As such, this approach, together with the clustering of functional

modules, can overcome limitations of knowledge. The biological relevance of these functional modules and hub genes cannot be surmised on the basis of associations with phenotypic traits and without credence to the molecular mechanisms that drove them. Co-expression links may arise from any source, from unresolved batch-effects to mRNA degradation patterns, and all sources of correlation are indistinguishable (Gaiteri, Ding et al. 2014). In the Illumina data set spurious module:trait associations may arise from the use of divergent tissue types (cartilage and tendon), or from batch-effects.

4.4.3: Consensus hubs associated with alginate culture systems

In spite of the caveats outlined above the most significant findings from consensus network analysis should be considered. This study was concerned with determining central regulators of the gene expression profiles found with three-dimensional cultures. No significant associations were found for tenocytes in fibrin cultures alone. For alginate cultures a consensus module was identified across the two data sets, which a) had a strong association with the alginate culture phenotype, and b) defined *Pi15*, *Chi3l*, *Serpina3n* and *Gpnmb* as potential core regulators. Of these *Pi15* and *Gpnmb* were both found to be more highly expressed in alginate cultures in both data sets by differential expression analysis. This does not preclude, however, involvement of the remaining. Notably, the model culture associated meta-module had a strong functional annotation relating to cell differentiation and development. It would be expected that genes with high intra-modular connectivity also have strong evidence of protein-protein interactions, however from STRING analysis there is little support for this statement for these limited data sets. It is, however, worth considering the relevance of these genes to cartilage function and differentiation.

Osteoactivin/transmembrane glycoprotein NMB in the protein encoded by *Gpnmb* and, as the name intimates, it has well-established relevance to osteoblast and osteoclast differentiation as shown by the work of Abdelmagid and colleagues (Selim, Abdelmagid et al. 2003, Abdelmagid, Barbe et al. 2008, Belcher, Rico et al. 2010) with evidence that this is BMP2 mediated (Abdelmagid, Barbe et al. 2007). It may be in secreted or transmembrane form (Abdelmagid, Barbe et al. 2008). Mutations in *Gpnmb* result in a loss of trabecular mass, reduced osteoid and mineralized regions and increased number of osteoblasts; there was enhanced levels of TGF-β receptors and phosphorylation of Smad-2/3 in mice with a *Gpnmb* loss-of-function mutation indicating a role for TGF-β signaling associated with osteoactivin (Abdelmagid, Belcher et al. 2014). Temporal localization of mRNA and protein expression in fracture repair coincided with chondrogenesis; staining was found in hypertrophic chondrocytes in the epiphyseal growth plates in normal bone, but also was increased in the fracture-healing callus (Abdelmagid, Barbe et al. 2010). In work by Karlsson, *et al* (2010) *Gpnmb* was found to be more highly expressed in OA-derived cartilage (approximately 9-fold change) than in cartilage from normal donors (Karlsson, Dehne et al. 2010). In micromass cultures derived from murine limb buds (E11.5) James, *et al* (2005) found higher expression of *Gpnmb* with Affymetrix microarrays between day 3 and day 15 of culture (James, Thomas et al. 2005). There is clear evidence to support a role for *Gpnmb* in the differentiation of cartilage and bone cells; whether it is a core regulator of chondrogenesis has not demonstrated and requires further validation of expression in alginate bead cultures.

Both *Pi15* and *Serpina3* code peptidase inhibitors and have been found to be differentially expressed in differentiating chondrocytes and diseased cartilage.

Serine peptidase inhibitor (Clade, A, member 3)/alpha-1-antichymotrypsin is encoded by Serpina3; the rat homolog is *Serpina3n*. Expression of *Serpina3* in differentiating MSCs and dedifferentiating chondrocytes has been documented (Boeuf, Steck et al. 2008). Peptidase inhibitor 15, *Pi15*, a trypsin inhibitor, was shown to be markedly suppressed in Grade I (Pritzker)-damaged cartilage in an monoiodoacetate-induced arthritis model, in addition to *Serpina3* (Nam, Perera et al. 2011).

These genes seem, ostensibly, rational targets for further investigation given their associations with cartilage phenotypes. Further investigation of the impact of gain- and loss- of function of these genes and their protein products on differentiation markers is warranted.

4.4.5: Data and study limitations

One of the key advantages of using the WGCNA approach over Bayesian network modeling of gene expression networks is that the latter requires very large sample sizes that are usually not available for microarray studies (Zhao, Langfelder et al. 2010). In this study it is apparent that even moderately sized microarray studies, such as the Affymetrix study (n=24), may not adequately meet the scale-free topology criterion required for this methodology. This is borne out in the indistinct gene cluster dendrogram for this dataset and the failure to adequately meet the approximate scale-free criterion. The Illumina data set (n=36), however, does approximate scale-free topology and had much more distinct module definitions.

In a study of murine chondrocyte differentiation Suwanwela, *et al* (2011) applied the WGCNA methodology to gene expression data from 27 microarrays, each

arising from the rib cartilage of a different recombinant inbred strain of mouse (Suwanwela, Farber et al. 2011). This study defined 14 modules, however, evident in the cluster dendrogram was the high similarity between all modules, modules were unusually small and comparable gene ontology annotations were evident between the modules. Quantitative findings associated with bone growth were correlated to module eigengenes, comparable to this study, and had significant associations at $p<0.05$. In the results presented in this analysis multiple narrowly significant associations are found between traits and modules, however, there is often poor correlation between the gene significance for a trait and module membership, which undermines any definition of a hub gene. It could be suggested that a higher level of significance is required in these studies than that often used as a standard threshold of significance. In this analysis for example, module-trait correlations less than 0.9 were considered to be tenuous associations.

The analysis presents strong associations between phenotypes and module eigengenes, however, the definition of these modules is derived from the co-expression of genes from disparate data sources. In the Illumina data set native cartilage and tendon are very distinct phenotypes and the strong association of module eigengenes with these phenotypes is not unexpected. It is probable, therefore, that the identification of modular hub genes arise uniquely within this dataset. For example, the identification of robust cartilage-associated genes as hubs for the 'cartilage' trait is not serendipitous, but likely to arise from the analysis of a small, divergent sample group. Furthermore, it only reinforces our differential expression analysis and fails to describe any emergent properties of the system. Consequently, results of hub gene analysis in single data sets are included here to demonstrate the outcomes and limitations of the methodology in small

data sets, rather than as an assertion that these genes are elusive core regulators. Those genes identified as core regulators of chondrocytes in alginate beads are relevant to chondrocyte differentiation status, but *Gpnmb*, for example, is highly expressed in both datasets and so may result in spurious correlations.

The co-expression analysis used here circumvents many of the issues associated with cross-platform microarray meta-analysis (see **Chapter 5**) by comparing network structures rather than absolute gene expression changes. The average rank gene expression was highly correlated between the data sets, which suggested that these sources were sufficiently comparable.

Whilst the benefits of comprehensively and inclusively defining the modular organization of a biological system are clear it is evident from this analysis that caution in extrapolating results from co-expression network analysis should be taken. The fallacy of correlation and causation can be readily exploited in under-powered studies with discrete phenotypes and as such findings would need to be challenged in a wider data context and mechanistic processes explored further.

4.4.5: Leucine zipper *Lzts2* silencing does not influence markers of chondrocyte differentiation in monolayer culture

As defined in **Chapter 1** the systems biology philosophy expects cyclical data gathering and analysis in response to targeted perturbations of the system under investigation. In exploration of the Illumina data using weighted gene network co-expression analysis the leucine zipper, putative tumour suppressor 2, *Lzts2*, was defined as a member of a module with strong association with the gene expression profiles in monolayer culture. Furthermore, it had high module membership and intra-modular connectivity. This would suggest that it had potential to be a central

regulator of this functional module. Higher expression of *Lzts2* in monolayer culture in both tenocytes and chondrocytes advocated further investigation.

Expression of *Lzts2* inhibits the nuclear translocation of beta-catenin (Thyssen, Li et al. 2006); in turn the transactivation of *Lzts2* is mediated by NF-κB activity. Reduction in the expression of *Lzts2* by RNA interference has been shown to promote the nuclear translocation of beta-catenin and NF-κB activity in adipose-derived human MSCs (Hyun Hwa, Hye Joon et al. 2008). These findings supported the complex interactions between Wnt- and NF-κB signaling. Both NF-κB and Wnt-signalling (through GSK3) are downstream effectors of the PI-3K/Akt-signalling cascade; GSK3, which phosphorylates β-catanin and targets it for degradation, has been shown to interact with *Lzts2* (Pilot-Storck, Chopin et al. 2010). Therefore, perturbation of a cross-pathway regulator, such as *Lzts2*, was considered to be a rational target.

Critical roles for Wnt-/beta-catenin signalling in chondrocyte development and differentiation is well recognised. Constitutive knock-out of beta-catenin is lethal to embryos at E7.5 prior to the formation of skeletal units (Haegel, Larue et al. 1995); enhanced chondrogenesis is observed in the conditional deletion of beta-catenin in mesenchymal condensations (Day, Guo et al. 2005, Hill, Später et al. 2005) and the inverse is true where beta-catenin fails to be degraded by GSK3 (Ryu, Kim et al. 2002); whilst delayed maturation is evident in growth-plate chondrocytes in a COL2A1-ICAT model that functionally inhibits beta-catenin (Chen, Zhu et al. 2008). Consequently, the implicated role of beta-catenin in the development of osteoarthritis is through the promotion of terminal differentiation in articular hyaline cartilage (Zhu, Tang et al. 2009, Wu, Zhu et al. 2010).

In a quantitative PCR study there was no evidence to support an effect of *Lzts2* inhibition on markers of differentiation status in dedifferentiated chondrocytes. Reasons for this may be technical or relate to redundancy within the system, i.e. knock-down does not sufficiently disturb the functional network. The genes that were defined as markers of differentiated, or de-differentiated status, may be only weakly discriminatory outside the context of the Illumina study in which they were defined.

Although the efficiency of transfection was high (>80%) the percentage knockdown was only 77% with a single siRNA to *Lzts2*. Additionally, in response to apparently cytotoxic effects on chondrocytes from extended incubation with the transfection reagent an exposure period of 6 hours was used. Although this in line with evidence from the manufacturer, this study could have benefitted from more robust inhibition, either through the use of two siRNA or a longer exposure period.

Finally, although isogenic tissue sources were used three biological replicates, and pooled technical replicates, are likely to be insufficient to resolve subtle changes in gene expression related to differentiation status. Primer efficiency was also variable, *Lzts2* was 90%, which could be considered poor for these profiling studies. Analysis using PCR arrays to a battery of differentiation and signalling pathway markers may reduce technical errors.

4.4.6: Closing statements

This study implemented WGCNA, a systems biology approach to objectively assess global transcriptome network properties between data sets arising from two different microarray platforms. This methodology supported findings from earlier

analysis by defining genes within modules that were strongly expressed in chondrocytes in alginate cultures and identified *Gpnmb* and *Pi15* as consensus network hubs. Knockdown of a NF-κB and Wnt-signalling cross-pathway regulator, *Lzts2*, failed to have an impact on the expression levels of markers of differentiation or dedifferentiation. The expression profiles arose from small data sets and related to highly divergent phenotypes and as such spurious correlations in the data may arise with no wider biological significance to functional modules and hub gene identification. In order to make more general statements about discrete chondrocyte and tenocyte responses to different conditions would require larger data sets or microarray meta-analysis and further validation of expression profiles of candidate hubs.

References

Abdelmagid, S., M. Barbe, I. Arango-Hisijara, T. Owen, S. Popoff and F. Safadi (2007). "Osteoactivin acts as downstream mediator of BMP-2 effects on osteoblast function." Journal of Cellular Physiology **210**(1): 26-37.

Abdelmagid, S., M. Barbe, M. Hadjiargyrou, T. Owen, R. Razmpour, S. Rehman, S. Popoff and F. Safadi (2010). "Temporal and spatial expression of osteoactivin during fracture repair." Journal of Cellular Biochemistry **111**(2): 295-309.

Abdelmagid, S., M. Barbe, M. Rico, S. Salihoglu, I. Arango-Hisijara, A. H. Selim, M. Anderson, T. Owen, S. Popoff and F. Safadi (2008). "Osteoactivin, an anabolic factor that regulates osteoblast differentiation and function." Experimental Cell Research **314**(13): 2334-2351.

Abdelmagid, S., J. Belcher, F. Moussa, S. Lababidi, G. Sondag, K. Novak, A. Sanyurah, N. Frara, R. Razmpour, F. Del Carpio-Cano and F. Safadi (2014). "Mutation in osteoactivin decreases bone formation in vivo and osteoblast differentiation in vitro." The American Journal of Pathology **184**(3): 697-713.

Albert, R. (2005). "Scale-free networks in cell biology." Journal of Cell Science **118**(21): 4947-4957.

Barabási, A.-L. and Z. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nature Reviews: Genetics **5**(2): 101-113.

Belcher, J., M. Rico, S. Ablemagid, A. Monroy, S. Popoff and F. Safadi (2010). OA/Gpnmb Mutation Impairs Osteoblast and Osteoclast Differentiation and Function. FASEB 2010 Experimental Biology Annual Meeting, Anaheim, CA, USA.

Boeuf, S., E. Steck, K. Pelttari, T. Hennig, A. Buneb, K. Benz, D. Witte, H. Sültmann, A. Poustka and W. Richter (2008). "Subtractive gene expression profiling of articular cartilage and mesenchymal stem cells: serpins as cartilage-relevant differentiation markers." Osteoarthritis and Cartilage **16**(1): 48-60.

Calabrese, G., B. Bennett, L. Orozco, H. Kang, E. Eskin, C. Dombret, O. De Backer, A. Lusis and C. Farber (2012). "Systems Genetic Analysis of Osteoblast-Lineage Cells." PLoS Genetics **8**(12): e1003150.

Chen, M., M. Zhu, H. Awad, T.-F. Li, T.-J. Sheu, B. Boyce, D. Chen and R. O'Keefe (2008). "Inhibition of beta-catenin signaling causes defects in postnatal cartilage development." Journal of Cell Science **121**(Pt 9): 1455-1465.

Cline, M., M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. Pico, A. Vailaya, P.-L. Wang, A. Adler, B. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. Warner, T. Ideker and G. Bader (2007). "Integration of biological networks and gene expression data using Cytoscape." Nature Protocols **2**(10): 2366-2382.

Conesa, A. and A. Mortazavi (2014). "The common ground of genomics and systems biology." BMC Systems Biology **8 Suppl 2**: S1.

Cosgrove, B. D., L. G. Griffith and D. A. Lauffenburger (2008). "Fusing Tissue Engineering and Systems Biology Toward Fulfilling Their Promise." Cellular and Molecular Bioengineering **1**(1): 33-41.

Day, T., X. Guo, L. Garrett-Beal and Y. Yang (2005). "Wnt/beta-catenin signaling in mesenchymal progenitors controls osteoblast and chondrocyte differentiation during vertebrate skeletogenesis." Developmental Cell **8**(5): 739-750.

Franceschini, A., D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering and L. Jensen (2013). "STRING v9.1: protein-protein interaction networks, with increased coverage and integration." Nucleic Acids Research **41**(Database issue): D808-D815.

Gaiteri, C., Y. Ding, B. French, G. C. Tseng and E. Sibille (2014). "Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders." Genes, Brain and Behavior **13**(1): 13-24.

Haegel, H., L. Larue, M. Ohsugi, L. Fedorov, K. Herrenknecht and R. Kemler (1995). "Lack of beta-catenin affects mouse development at gastrulation." Development (Cambridge, England) **121**(11): 3529-3537.

Hartwell, L. H., J. J. Hopfield, S. Leibler and A. W. Murray (1999). "From molecular to modular cell biology." Nature **402**(6761 Suppl): C47-C52.

Hawrylycz, M., E. Lein, A. Guillozet-Bongaarts, E. Shen, L. Ng, J. Miller, L. van de Lagemaat, K. Smith, A. Ebbert, Z. Riley, C. Abajian, C. Beckmann, A. Bernard, D. Bertagnolli, A. Boe, P. Cartagena, M. Chakravarty, M. Chapin, J. Chong, R. Dalley, B. Daly, C. Dang, S. Datta, N. Dee, T. Dolbeare, V. Faber, D. Feng, D. Fowler, J. Goldy, B. Gregor, Z. Haradon, D. Haynor, J. Hohmann, S. Horvath, R. Howard, A. Jeromin, J. Jochim, M. Kinnunen, C. Lau, E. Lazarz, C. Lee, T. Lemon, L. Li, Y. Li, J. Morris, C. Overly, P. Parker, S. Parry, M. Reding, J. Royall, J. Schulkin, P. Sequeira, C. Slaughterbeck, S. Smith, A. Sodt, S. Sunkin, B. Swanson, M. Vawter, D. Williams, P. Wohnoutka, R. Zielke, D. Geschwind, P. Hof, S. Smith, C. Koch, S. Grant and A. Jones (2012). "An anatomically comprehensive atlas of the adult human brain transcriptome." Nature **489**(7416): 391-399.

Hill, T., D. Später, M. Taketo, W. Birchmeier and C. Hartmann (2005). "Canonical Wnt/beta-catenin signaling prevents osteoblasts from differentiating into chondrocytes." Developmental Cell **8**(5): 727-738.

Horvath, S. (2011). Adjacency Functions and Their Topological Effects. Weighted Network Analysis, Springer New York**:** 77-89.

Horvath, S. (2011). Clustering Procedures and Module Detection. Weighted Network Analysis, Springer New York**:** 179-206.

Hyun Hwa, C., J. Hye Joon, S. Ji Sun, B. Yong Chan and J. Jin Sup (2008). "Crossregulation of beta-catenin/Tcf pathway by NF-kappaB is mediated by lzts2 in human adipose tissue-derived mesenchymal stem cells." Biochimica et Biophysica Acta **1783**(3): 419-428.

James, C., Thomas, V. Ulici, M. Underhill and F. Beier (2005). "Microarray Analyses of Gene Expression during Chondrocyte Differentiation Identifies

Novel Regulators of Hypertrophy." Molecular Biology of the Cell **16**(11): 5316-5333.

Karlsson, C., T. Dehne, A. Lindahl, M. Brittberg, A. Pruss, M. Sittinger and J. Ringe (2010). "Genome-wide expression profiling reveals new candidate genes associated with osteoarthritis." Osteoarthritis and Cartilage **18**(4): 581-592.

Killcoyne, S., G. Carter, J. Smith and J. Boyle (2009). "Cytoscape: a community-based framework for network modeling." Methods in Molecular Biology (Clifton, N.J.) **563**: 219-239.

Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. Jones and M. Marra (2009). "Circos: An information aesthetic for comparative genomics." Genome Research **19**(9): 1639-1645.

Langfelder, P. and S. Horvath (2007). "Eigengene networks for studying the relationships between co-expression modules." BMC Systems Biology **1**(1): 54.

Langfelder, P. and S. Horvath (2008). "WGCNA: an R package for weighted correlation network analysis." BMC Bioinformatics **9**(1): 559.

Langfelder, P. and S. Horvath (2012). "Fast R Functions for Robust Correlations and Hierarchical Clustering." Journal of Statistical Software **46**(11): i11.

Langfelder, P., P. Mischel and S. Horvath (2013). "When Is Hub Gene Selection Better than Standard Meta-Analysis?" PLoS ONE **8**(4): e61505.

Langfelder, P., B. Zhang and S. Horvath (2008). "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R." Bioinformatics (Oxford, England) **24**(5): 719-720.

Miller, J., S. Horvath and D. Geschwind (2010). "Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways." Proceedings of the National Academy of Sciences of the United States of America **107**(28): 12698-12703.

Nam, J., P. Perera, J. Liu, B. Rath, J. Deschner, R. Gassner, T. Butterfield and S. Agarwal (2011). "Sequential alterations in catabolic and anabolic gene expression

parallel pathological changes during progression of monoiodoacetate-induced arthritis." PloS ONE **6**(9): e24320.

Pilot-Storck, F., E. Chopin, J.-F. Rual, A. Baudot, P. Dobrokhotov, M. Robinson-Rechavi, C. Brun, M. Cusick, D. Hill, L. Schaeffer, M. Vidal and E. Goillot (2010). "Interactome mapping of the phosphatidylinositol 3-kinase-mammalian target of rapamycin pathway identifies deformed epidermal autoregulatory factor-1 as a new glycogen synthase kinase-3 interactor." Molecular & Cellular Proteomics **9**(7): 1578-1593.

Rajagopalan, P., S. Kasif and T. M. Murali (2013). "Systems Biology Characterization of Engineered Tissues." Annual Review of Biomedical Engineering **15**(1): 55-70.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai and A. L. Barabási (2002). "Hierarchical organization of modularity in metabolic networks." Science (New York, N.Y.) **297**(5586): 1551-1555.

Ryu, J.-H., S.-J. Kim, S.-H. Kim, C.-D. Oh, S.-G. Hwang, C.-H. Chun, S.-H. Oh, J.-K. Seong, T.-L. Huh and J.-S. Chun (2002). "Regulation of the chondrocyte phenotype by beta-catenin." Development (Cambridge, England) **129**(23): 5541-5550.

Selim, A., S. Abdelmagid, R. Kanaan, S. Smock, T. Owen, S. Popoff and F. Safadi (2003). "Anti-osteoactivin antibody inhibits osteoblast differentiation and function in vitro." Critical Reviews in Eukaryotic Gene Expression **13**(2-4): 265-275.

Slonim, D. (2002). "From patterns to pathways: gene expression data analysis comes of age." Nature Genetics **32 Suppl**: 502-508.

Suwanwela, J., C. Farber, B.-L. Haung, B. Song, C. Pan, K. Lyons and A. Lusis (2011). "Systems genetics analysis of mouse chondrocyte differentiation." Journal of Bone and Mineral Research **26**(4): 747-760.

Thyssen, G., T.-H. Li, L. Lehmann, M. Zhuo, M. Sharma and Z. Sun (2006). "LZTS2 is a novel beta-catenin-interacting protein and regulates the nuclear export of beta-catenin." Molecular and Cellular Biology **26**(23): 8857-8867.

Wang, Y., D. J. Miller and R. Clarke (2008). "Approaches to working in high-dimensional data spaces: gene expression microarrays." British Journal of Cancer **98**(6): 1023-1028.

Wu, Q., M. Zhu, R. Rosier, M. Zuscik, R. O'Keefe and D. Chen (2010). "Beta-catenin, cartilage, and osteoarthritis." Annals of the New York Academy of Sciences **1192**(1): 344-350.

Yang, Y., L. Han, Y. Yuan, J. Li, N. Hei and H. Liang (2014). "Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types." Nature Communications **5**: Article number: 3231.

Yip, A. and S. Horvath (2007). "Gene network interconnectedness and the generalized topological overlap measure." BMC Bioinformatics **8**(1): 22.

Zhang, B. and S. Horvath (2005). "A general framework for weighted gene co-expression network analysis." Statistical Applications in Genetics and Molecular Biology **4**(1): Article 17.

Zhao, W., P. Langfelder, T. Fuller, J. Dong, A. Li and S. Hovarth (2010). "Weighted Gene Coexpression Network Analysis: State of the Art." Journal of Biopharmaceutical Statistics **20**(2): 281-300.

Zhu, M., D. Tang, Q. Wu, S. Hao, M. Chen, C. Xie, R. Rosier, R. O'Keefe, M. Zuscik and D. Chen (2009). "Activation of beta-catenin signaling in articular chondrocytes leads to osteoarthritis-like phenotype in adult beta-catenin conditional activation mice." Journal of Bone and Mineral Research **24**(1): 12-21.

Zhu, X., M. Gerstein and M. Snyder (2007). "Getting connected: analysis and principles of biological networks." Genes & Development **21**(9): 1010-1024.

A Systems Biology Approach To

Musculoskeletal Tissue Engineering:

Transcriptomic And Proteomic Analysis Of

Cartilage And Tendon Cells

By

Alan James Mueller

Volume II

Chapters 5 - 8

5: Integration of Cartilage and Tendon Microarray Expression Profiles Reveals Cross-species Preservation of Gene Modules

"*A big computer, a complex algorithm and a long time does not equal science.*"

Robert Gentleman

Abstract

Integration and comparison of gene expression profiles across multiple studies and microarray platforms is fraught with difficulties, which can limit the biological relevance of findings. Microarray studies are often underpowered and in musculoskeletal biology limited biological replicates arising from rare samples compound this problem. The integration of data from multiple small studies can improve the statistical power of an analysis. The use of rodent *in vivo* models of musculoskeletal disease, or as part of *in vitro* studies, is frequent, however, the translational relevance to complex and diverse human disease is not always clear. Furthermore, three-dimensional culture systems are used to mimic spatial conditions for chondrocytes and tenocytes, but these methods have not been systematically considered relative to normal or diseased tissue.

The relative merits of data-merging approaches, including cross-platform normalisation, and sequence- and target-matching of oligo probes across platforms, were investigated to integrate expression data into single species-specific meta-matrices.

This study presents the first cross-species systematic integration of cartilage and tendon gene expression data utilising a data-merging approach in the rat and human. A global transcriptomic network comparison approach was employed applying gene co-expression analysis, WGCNA, to define functional modules associated with disease and three-dimensional culture phenotypes. By comparing the networks structures across species the suitability of rodent models to inform human disease is considered.

An *IL-6* containing module was found to be significantly associated with a perturbed chondrocyte phenotype and, more specifically, with alginate bead cultures. From this module thirteen, mostly pro-inflammatory, genes were found, by a nearest shrunken centroids approach, to adequately discriminate between three-dimensional culture systems and all other samples. This was consistent across rat and human studies and when compared to an independent data set

5.1: Introduction

Increasingly gene expression profiling data is submitted to public-access repositories, including Gene Expression Omnibus (Barrett, Troup et al. 2011) and ArrayExpress (Rustici, Kolesnikov et al. 2013), and is available for re-use by researchers in large-scale integrative analyses. Relative to other areas of research there are few transcriptomic profiles of normal musculoskeletal tissues available in these repositories; the small number of biological replicates in each experiment can represent an important bottleneck in the exploration of biological problems. To improve the statistical power and return more reliable conclusions from transcriptome profiling some form of integrative analysis is required to increase the number of samples (Taminau, Lazar et al. 2014).

5.1.1: Systematic analysis of microarrays

The terminology associated with the re-use of gene expression data is inconsistent and requires some explanation. Commonly the term 'meta-analysis' has referred to the systematic and standardised study of a focused issue in research literature. In the main a meta-analysis considers a particular intervention, usually medical, across all the published literature allowing both quantitative and qualitative assessment of the relative strength of each of the individual studies (Russo 2007). Combining the marginal findings in small, underpowered studies may facilitate the detection of statistically significant responses to an intervention; equally, a meta-analysis may repudiate weak findings in individual studies.

Integral to the development of a meta-analysis, as described by Russo (2007) are the inclusion of: i) an explicit study question; ii) an exhaustive literature search; iii)

defined criteria for inclusion of a study, or data abstraction, which may include sourcing raw data, quality scoring studies and the presentation of excluded data; iv) and the description of statistical tests, in particular, tests for sample homogeneity, and the use of fixed or random effects models (Russo 2007).

Recently, the term 'meta-analysis' has been applied to any systematic study of multiple smaller studies including genome-wide association studies (GWAS) (Evangelou and Ioannidis 2013) and the analysis of microarray data sets (Rudy and Valafar 2011). The availability of microarray data from public-access repositories has encouraged researchers to integrate multiple microarray studies, although these investigations may lack the stringency with which clinical meta-analysis have been performed. An explicit study question may not be apparent at the data collection stage, for example GWAS meta-analysis considers the presence of any genetic risk loci for a population. The general structure of phenotype specification, data collection, data inclusion/exclusion, and reduction of data heterogeneity still apply, however, to these other forms of meta-analyses.

Whilst some have used 'meta-analysis' for any large-scale analysis of microarray expression data in this study the term will be restricted to discussions relating to the integration at the interpretive level, and 'data-merging' to describe the agglomeration of raw expression data from multiple data sets, consistent other investigators (Sarmah and Samarasinghe 2010, Taminau, Lazar et al. 2014). These definitions are explored further below.

Motivations for microarray data integration

There are several principal drivers for the integration of microarray expression data. Integration, a) increases sample sizes, thereby improving statistical power; b)

facilitates a broader understanding of a biological problem, c) compensates for data errors and missing data in individual studies, d) provides more accurate and consistent data mining, e) explores the variance and noise in the data, and f) can define biological markers and prognostic signatures not evident in small analyses (Sarmah and Samarasinghe 2010). Furthermore, consistent with the tenets of reduction, refinement and replacement, *in silico* analysis allows the re-use of existing data to derive novel results and inform future experimental design.

Issues of statistical power, for example, often arise in microarray studies as too few biological replicates are used, these are often not independent samples (i.e. biological replicates are re-used within a study), and they measure the expression levels of large number of genes simultaneously (Ramasamy, Mondry et al. 2008). Through gathering data from multiple sources there is potential to discover of emergent properties of the data, explores new biological insights and improve the statistical power of gene expression analysis (Taminau, Meganck et al. 2012).

Approaches to expression data integration

There is no definitive methodology for tackling microarray data integration with studies employing a number of strategies. Data is rarely considered at the probe-level, with strategies considering higher order analysis, for example, published differentially expressed gene lists, combining *p*-values, combining ranks or effect sizes (Tseng, Ghosh et al. 2012). An essential step in approaching an over-arching analysis of microarray data is to ensure that there is a consistent and standardized data handling. This can mean the application of the same pre-processing and normalization techniques, incorporating an adjustment for batch effects and platform, and maximising consensus gene identifiers across platforms (Taminau,

Meganck et al. 2012). Whilst all this is done the true biological variance that exists in the individual datasets has to be retained.

Rung and Brazma (2012) present sage advice on the pitfalls of tackling such microarray meta-analysis, in particular that 'summary level meta-analysis is often a better option' (Rung and Brazma 2012). Whether the analysis is based upon the summary results from individual experiments or based upon the merging of raw data is a choice made based upon what will be most informative to the researcher. The intended question may be unavoidably and strongly biased by the differences that are found in diverse experimental conditions.

These methods usually lead to four main types of microarray meta-analysis, as described by Tseng, *et al* (2012): i) differential gene expression; ii) pathway analysis; iii) co-expression analysis; iv) prediction analysis. Each resolves the data in a different way and produces results with varied potential of generalization or clinical utility. The integration and meta-analysis of microarray studies by gene co-expression network analysis has facilitated and yielded outputs that provide a functional appreciation of the tissues under investigation (Miller, Horvath et al. 2010, Hawrylycz, Lein et al. 2012). Methods already explored in **Chapter 4** are re-visited in this study to facilitate integration of multiple gene expression studies.

In essence the re-use of expression data is a form of integrative analysis that may involve either, a) integration at the interpretive level, or b) integration at the level of expression values (Sarmah and Samarasinghe 2010).

Integration at the interpretative level – 'meta-analysis'

The approach, which may be termed 'meta-analysis', requires the comprehensive and consistent re-analysis of each data set followed by the combination of

summary statistics (*P*-values, ranks) at an interpretative level. Consistency and comparability in the raw data collection technique, whether that is patient-level data, microarray expression studies or RNA-sequencing data, is critical. In the case of microarrays, where the raw data arises from different technology platforms and platform generations, neither consistency nor comparability can be assumed and, as such, these types of studies are more correctly termed 'cross-platform' studies (Rudy and Valafar 2011).

Data-merging – integration at the level of expression values

Microarray data-merging utilises raw expression data from multiple sources then transforms and normalises the data to make it numerically comparable. Many of the key issues relating to meta-analysis of microarrays are applicable to data-merging (Sarmah and Samarasinghe 2010, Taminau, Lazar et al. 2014). The output is a new, larger data set upon which further analysis is performed, for example, evaluation of prognostic gene signatures indicating suitability of therapeutic interventions (Xu, Tan et al. 2008). Data handling approaches are defined further below. Which approach is more consistent in terms of defining prognostic markers or differentially expressed genes is not clear in the literature, however, recent investigation suggests that significantly more differentially expressed genes may be found through a data-merging approach (Taminau, Lazar et al. 2014).

Obstacles to integration of expression studies

Whilst the motivation to integrate gene expression data is attractive a number of data-handling challenges retard easy implementation. Direct combination of gene expression studies is not possible without further normalisation, neither are independent studies from different microarray platforms directly comparable. Furthermore, in terms of experimental design, integration needs to be biologically

meaningful, rather than exhaustive harvesting and collation of incompatible samples (Sarmah and Samarasinghe 2010).

Concerns with regard to microarray data integration have been evident for over a decade and despite the implementation of minimal expected standards (MIAME) (Brazma, Hingamp et al. 2001, Moreau, Aerts et al. 2003) for microarray data deposited in public databases one study found little over a third of data submitted met the format required and quality standards (Larsson and Sandberg 2006). The absence of raw data (permitting future re-annotation and application of novel computational methods) and low quality noisy data, were highlighted as particular hurdles to integration.

Microarray data integration is obstructed by relevant studies having been undertaken on microarray platforms from numerous vendors, some of which may be archaic and poorly annotated. In addition to the biological variation inherent in each study sample microarray platforms each have unique chemistries, hybridization protocols, probe identifiers and probe lengths. Furthermore, the well-described issues associated with systematic bias, relating to 'batch effects' associated with temporal and technical variations (Leek, Scharpf et al. 2010, Chen, Grennan et al. 2011), are unlikely to be considered in a microarray meta-analysis or cross-platform data-merging as information relating to the time and location of microarray processing are often not available. Noise is an inescapable problem in any microarray study, but this is propagated further with the addition of data from diversified sources (Sarmah and Samarasinghe 2010).

Ramasamy, *et al* (2008) provided a practical framework by which to tackle microarray meta-analysis with the greatest emphasis being on the data collection

and curation methods (Ramasamy, Mondry et al. 2008). In choosing which data to work with feature-level/probe-level data was recommended, thereby allowing the researcher flexibility in which pre-processing and normalisation techniques were employed, and ensuring that the same method was used across all studies.

Dealing with obstacles to data-merging

Probe-matching

Two microarrays from two platforms can, in reality, never be fundamentally comparable. Platforms differ not only in their chemistries but also their probe designs and lengths. The design of the probes reflects a trade-off between sensitivity and specificity, the probe sequences defining the hybridization characteristics of the probes; as two platforms may not share probe sequences the data cannot be directly comparable (Rudy and Valafar 2011). Direct probe-level data integration would only be possible between datasets using the same or technically similar platforms, e.g. platform generations from the same manufacturer (Tseng, Ghosh et al. 2012); there is evidence that sequence matching of oligonucleotide probes across platforms would improve the reproducibility and concordance of a cross-platform analysis (Mecham, Klus et al. 2004, Carter, Eklund et al. 2005, Kuo, Liu et al. 2006, Ramasamy, Mondry et al. 2008, Allen, Wang et al. 2012). However, two of the studies above considered only comparisons between two platforms at a time (Allen (2012), Carter (2006)), whereas Kuo, *et al* (2006) considered ten platforms. The latter presented good correlation of expression measurements for exon-matched probes even where the probe sequences did not overlap, however, they could only find four genes for which probes from all ten platforms mapped to the same exon.

Normalisation strategies

Integration across microarray data sets often requires recursive data normalisation, for example, between samples in a data set and then across data sets; this can occur at the expense of the true biological variation in the data resulting in an 'over-smoothing' making any analysis largely redundant or introduces spurious correlations (Sîrbu, Ruskin et al. 2010). Subsequently correlations between interacting genes may be lost making the use of inferential algorithms downstream problematic; alternatively, as discussed in the previous chapter, such interventions may introduce spurious correlations between the data sets (Gaiteri, Ding et al. 2014). The complexity of rescaling approaches to gene expression data varies from *z-score* standardization (Cheadle, Cho-Chung et al. 2003a) to Bayesian batch-correction methods (Johnson, Li et al. 2007).

Sîrbu, *et al* (2010) collected arrays from three different platforms and, following three types of preprocessing on the raw data, performed cross-platform normalisation undertaken by i) *z*-score standardization (scaling such that expression values lay in [0,1], ii) a Bayesian approach to batch-effect removal (ComBat), and iii) an iterative *k*-means clustering technique (XPN). In this study it was found that for normalisation of time-series data for quantitative model inference a combination of loess normalisation during pre-processing and subsequent use of the XPN algorithm resulted in acceptably low variation.

Rudy and Valafar (2011) also found the XPN algorithm generated the highest inter-platform concordance, but only when treatment groups were of equal size (Rudy and Valafar 2011). In analysis of four methods for cross-platform normalisation they found that the DWD (Distance Weighted Discrimination) algorithm (Benito, Parker et al. 2004) was more robust to these inequalities. From

these observations there is clearly discordance between the capabilities of normalisation algorithms, their complexity, and the practicalities of sourcing groups of equal size, on the same platform, or establishing comparable co-variants and samples derived from the same experimental methodology.

5.1.2: Meta-analysis of cartilage and tendon pathologies

As highlighted in **Chapter 2**, the scope of gene expression studies of cartilage and tendon have been limited with little evaluation of the baseline expression profiles of the constituent cells nor further evaluation of their expression profiles in perturbed environments. Consequently there is, to the author's knowledge, no systematic integration of microarray data sets interrogating cartilage or tendon gene expression. Furthermore, there is no coherent or comprehensive understanding of the chondrocyte or tenocyte response profiles to perturbations on a global transcriptomic scale.

Of the meta-analyses that have been performed for cartilage and tendon pathologies the majority are clinical, intervention-based, investigations. Of the others genome-wide association studies (GWAS) of arthritic and rheumatic conditions are most prevalent. Microarray meta-analysis of osteoarthritis and rheumatoid arthritis invariably study either synovial tissue or use grossly normal and abnormal articular cartilage samples from the same patient. Given the numerous hurdles described it is not unexpected that systematic analysis of gene expression profiles is not common. For tendinopathies the lack of systematic analyses is more acute – in this review there were no non-clinical publications exploring gene expression of tenocytes across publications.

A number of GWAS studies have implicated various genetic risk loci in the development of osteoarthritis, including *GDF5* and *SMAD3* (Valdes, Spector et al. 2010, Reynard and Loughlin 2013), or found conflicting associations with other candidates, such as *IL-6* (Valdes, Arden et al. 2010, Cai, Sun et al. 2014). However, a recent meta-analysis of GWAS studies, considering 199 candidate genes, found that SNPs associated with only two genes (*COL11A1* and *VEGF*) had a statistical association with the development of osteoarthritis (Rodriguez-Fontenla, Calaza et al. 2014).

Animal models of musculoskeletal disease

For animals models to inform human musculoskeletal disease those proxies must be appropriate and relevant, the former referring to the comparability of a disease process in the model to that in the human, the latter considering the relevance of a complex whole animal model where simpler experimental set-ups would suffice (Pritzker 1994). This is problematic in the study of musculoskeletal disease where small studies, diverse models, and animal differing strains have been used for gene expression profiling. Interrogation of expression profiling studies is required to define whether regulatory mechanisms are present or absent in models of osteoarthritis relative to the complex and diverse human disease (Goldring 2012).

5.1.3: Summary and study rationale

This study is driven by three main concerns arising from the literature. First, there is no evidence of a systematic analysis of global transcriptomic networks having been undertaken to explore common regulatory mechanisms between rodent and human musculoskeletal disease or to validate these models. Second, *in-vivo* culture models have not been defined relative to diseased tissue, the assumption being made that their gene expression profiles are sufficiently comparable to normal

samples. Third, by reviewing the available studies gaps in community knowledge may be highlighted; this may facilitate development of future targeted research projects. It is prescient, therefore, that these concerns are challenged.

This study utilizes a data-merging approach to integrate comprehensive expression data from multiple small cartilage and tendon transcriptome profiling studies from the rat and human. A global transcriptome network comparison approach, WGCNA, is applied to define functional modules associated with disease and three-dimensional culture phenotypes. By comparing the networks structures across species the suitability of rodent models for informing human disease is considered.

To accommodate diverse data sources from relatively few studies a number of data integration methods are explored. An *IL-6* containing module is found to be significantly associated with a perturbed chondrocyte phenotype and, more specifically, with alginate bead cultures. From this module thirteen, mostly pro-inflammatory, genes were found, by a nearest shrunken centroids approach, to adequately discriminate between three-dimensional culture systems and all other samples. This was consistent across rat and human studies and when compared to an independent data set.

5.2: Methods

5.2.1: Identification of datasets and inclusion criteria

Curated, open-access microarray repositories ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) (Rustici, Kolesnikov et al. 2013) and Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) (Barrett, Troup et al. 2011) were trawled for data sets profiling tissues of musculoskeletal origin derived from *Rattus norvegicus*, namely: 'cartilage', 'tendon', and 'ligament', **Figure 5.1a**. Additionally, disease-orientated searches of data were undertaken using the terms: 'osteoarthritis', 'rheumatoid arthritis', 'tendinopathy', and 'tendinitis'. Further to this data sets were collected if they contained musculoskeletal developmental stages or directed differentiation studies, e.g. mesenchymal stem cells. All recorded studies available prior to June 2014 were considered. A literature search of PubMed (http://www.ncbi.nlm.nih.gov) was undertaken using the aforementioned terms to collect studies where expression data was not publically available and corresponding authors were contacted. All collected data sets were tested against the inclusion criteria described in **Table 5.1**. The signalment of the animal specimens (age, sex, breed) was recorded, but this did not influence inclusion of data in the study.

5.2.2 Data preparation

The microarray pre-processing and merging pipeline is provided in **Figure 5.1b**. Briefly, raw data was imported into *R* and arrays were visually assessed for systematic technical issues and underwent the same quality control appraisal as applied to microarray data in **Chapter 2** and **3**. Expression data was background corrected using the RMA algorithm (Irizarry, Hobbs et al. 2003) and a cyclic loess

normalisation method applied across each study data set, as described before. Probe sets were re-annotated with the appropriate Entrez gene identifier. Expression data for each gene was aggregated and collapsed into a single gene measurement consisting of the maximum mean expression value using the 'collapseRows' function in the `WGCNA` package (Miller, Cai et al. 2011). The output of this workflow was a normalised matrix of expression values consisting of one summarized gene per row. Each platform expression matrix had a differing number of genes at this stage. Data sets were intersected by Entrez gene identifiers for cross-study normalisation such that all studies contained the same gene identifiers. The matrix of aggregated data sets was termed a 'meta-matrix'. The final rat meta-matrix contained 170 microarray samples and profiled 11,152 genes derived from four Affymetrix platforms and was the input for weighted gene co-expression network analysis.

Inclusion criteria	**Exclusion Criteria**
Raw, probe-level data files were available and studies achieved >3 of the requirements of the MIAME conditions.	Some level of pre-processing had been undertaken, e.g. quantile normalisation.
Data sets were derived exclusively from *Rattus norvegicus*	Individual arrays within a dataset failed to meet quality control thresholds
Probe annotation files were available	Platforms with small transcript coverage, e.g. Affymetrix Rat U34A array
Datasets could be manipulated using the *R* programming platform	Data sets were re-used across multiple studies
Study design and sample phenotypes and descriptors were available.	Single biological replicate used over multiple arrays in a study.
Studies investigated cartilage, tendon or ligament in their native state or their cellular components *in vitro*.	Studies with fewer than three arrays in a study group.

Table 5.1: Definition of inclusion and exclusion criteria for microarray meta-analysis.

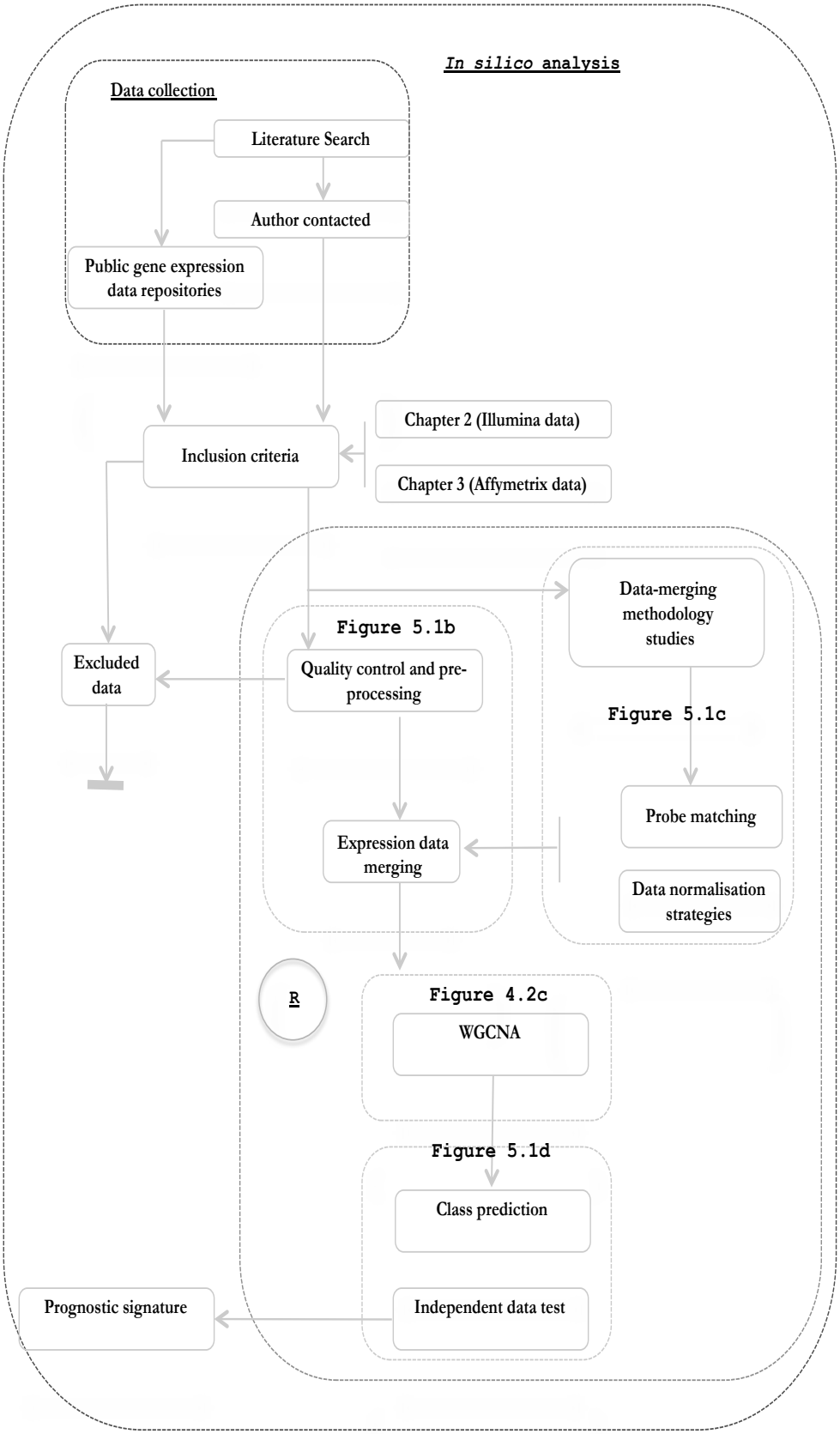


Figure 5.1a: Overview of experimental workflow for results presented in **Chapter 5**. Input data is derived from **Chapters 2** and **3**, public repositories and personal requests. Inclusion and exclusion criteria are defined in **Table 5.1**. Parallel studies informing and refining data analysis pipelines are defined in **Figure 5.1b – d** and WGCNA pipeline in **Figure 4.2c**.

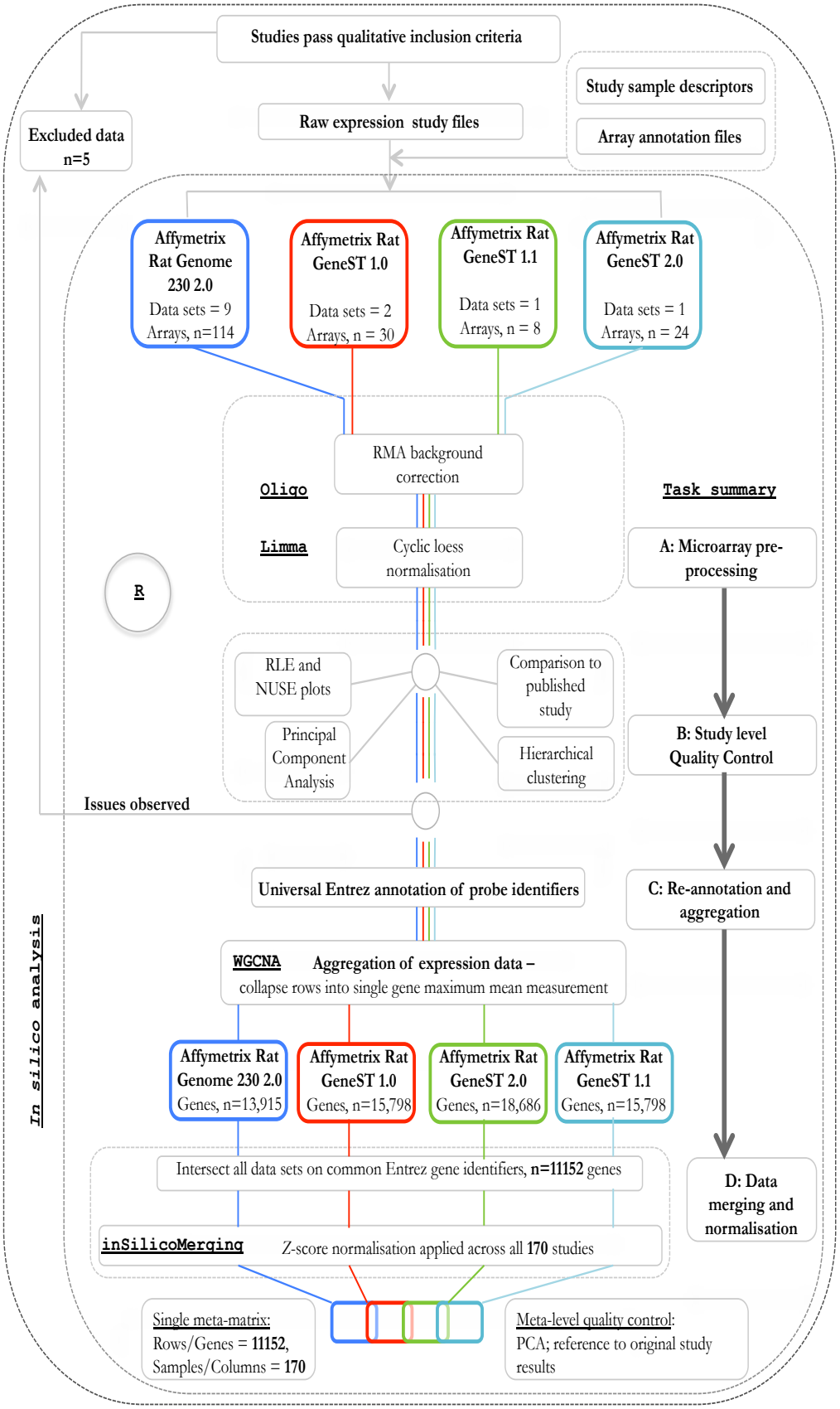


Figure 5.1b: Data merging pipeline for rat studies profiling chondrocyte and tenocyte transcriptomes from four Affymetrix microarray platforms. All analysis was undertaken in the R environment with critical work packages indicated, e.g. **`limma`**. All study data sets underwent separate analysis through tasks **A** to **D** before final merging and global normalisation. This process was repeated for human microarray data sets (n = 166).

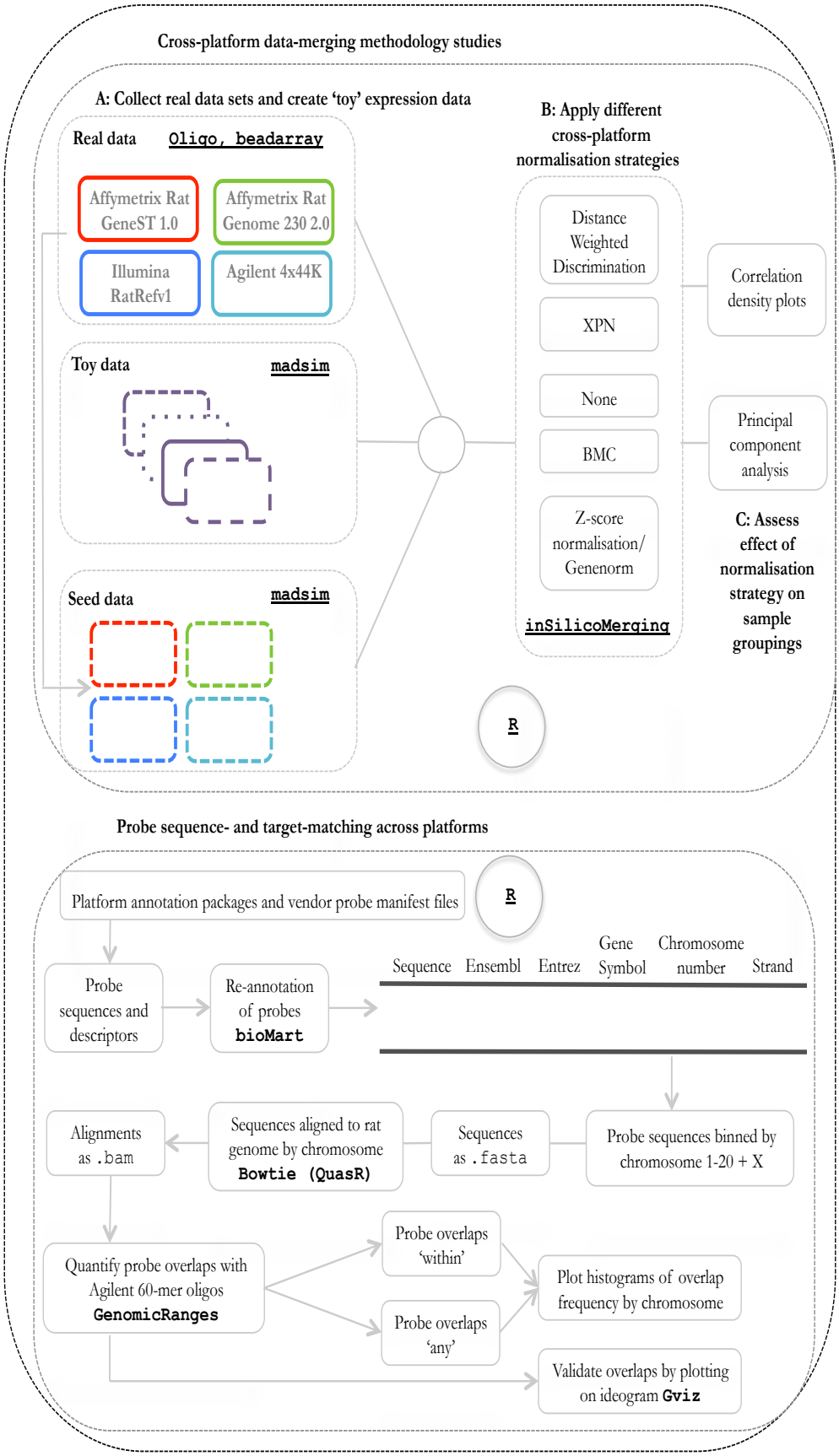


Figure 5.1c: Data manipulation and analysis pipelines for cross-platform normalisation (upper panel) and probe sequence- and target-matching across platforms (lower panel). Outcomes of studies informed **Figure 5.1b**. Critical work-packages within the R environment are provided, e.g. **`Gviz`**.

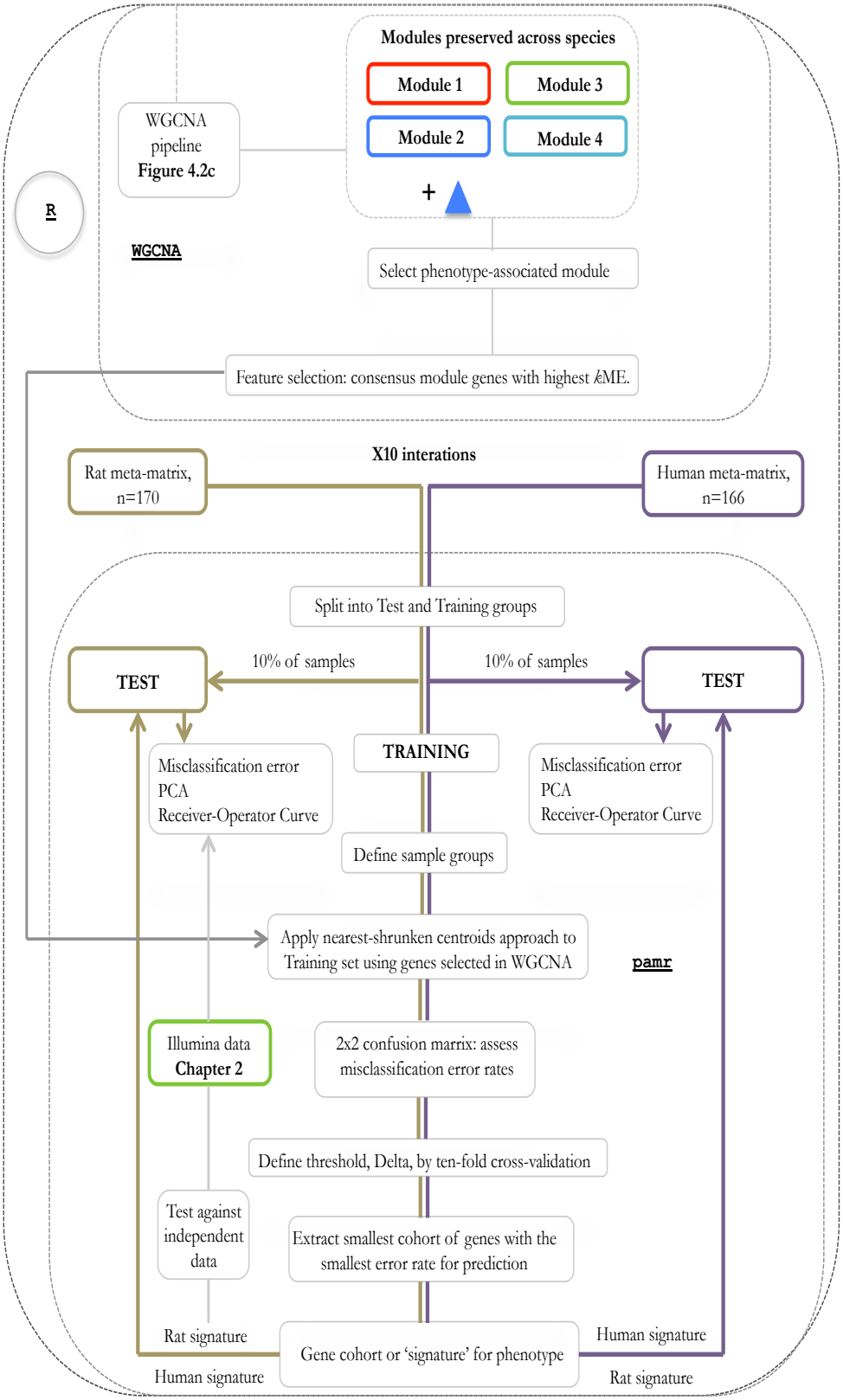


Figure 5.1d: Data manipulation and analysis pipelines for class-prediction using **pamr**. Genes were filtered for 'feature-selection' by restricting available genes to those found within a highly phenotype-associated consensus module in WGCNA. This was common to both human and rat analysis. Gene signatures identified for each species were used to classify test data from the same species and *vice versa*. Misclassification error rates and areas under receiver-operator curves defined the predictive success of each signature. Illumina expression data (**Chapter 2**) was used as an independent data data set against which to assess the rat signature.

5.2.3: Analysis of cross-platform normalisation techniques

Rat expression data was derived from four platforms from three manufacturers: Affymetrix Rat Genome 230 2.0, Illumina RatRef.v1, Affymetrix Gene ST 1.0, and Agilent 4x44k. In order to explore the impact of cross-platform normalisation strategies on data a series of test datasets were prepared, **Figure 5.1c upper panel**: i) simulated or 'toy' data set was prepared consisting of a set of four randomly generated arrays using the `madsim` *R* package (Dembélé 2013); ii) 'Seed' data sets were simulated arrays prepared as for (i), but used real data sets as the kernel around which the distribution of expression values was built. Both 'case' and 'control' data sets were produced; iii) real data sets from public repositories were chosen to represent each of the four platforms collected in this study. As many real data sets are underpowered and have few biological replicates three case and three controls were taken from each platform. Data sets used: Illumina (**Chapter 2**); Affymetrix 1.0 (Nam, Perera et al. 2011); Affymetrix 230 (Appleton, Pitelka et al. 2007); Agilent 44k (Zhang, Fang et al. 2012).

Several normalisation strategies available within the `inSilicoMerging` *R* package (Taminau, Meganck et al. 2012) were applied to all data sets: 'DWD' (Benito, Parker et al. 2004), 'XPN' (Shabalin, Tjelmeland et al. 2008), 'BMC' (Sims, Smethurst et al. 2008), 'Genenorm' (z-score normalisation (Cheadle, Cho-Chung et al. 2003a, Cheadle, Vawter et al. 2003b)) or 'NONE' (merging with no normalisation). Empirical-Bayes batch-correction method, 'COMBAT' (Johnson, Li et al. 2007), was not used as this method assumes covariates to be consistent across samples. Data for simulated and real data analysis were pre-processed and normalised in the same manner.

Single manufacturer data sets

The use of data sets arising from one platform provider only, Affymetrix, was considered. These are the most numerous for rat studies in the public repositories and all use 25-mer oligo 'probe-sets'. Three Affymetrix platform generations were represented in the cartilage and tendon data collected: Genome 230 2.0, GeneST1.0 and GeneST2.0. Data was pre-processed as described previously. One data set for each platform was considered initially (n=37).

5.2.4: Sequence- and target-matching of probes from three platforms

In order to maximize the number of available cartilage and tendon studies the most commonly used rat microarray platforms were assessed for probe sequence- and target-matching: i) Illumina RatRef.v1 BeadChip; ii) Affymetrix GeneST 1.0; iii) Affymetrix 230 Expression Array; iii) Agilent 4x44K version G4131F, **Figure 5.1c lower panel**. Unique probe and probe-set identifiers and their associated sequences were collected from Bioconductor annotation packages or from manifest files from the manufacturer or GEO.

Probes were re-annotated using functions within the `biomaRt` package, an interface with the Biomart database (www.biomart.org), implemented in *R* (Durinck, Moreau et al. 2005, Durinck, Spellman et al. 2009, Kasprzyk 2011). Probes were annotated with Ensembl and Entrez identifiers, common gene symbols, chromosome number and location (band/strand) for orientation purposes. Annotated probe sequence files were binned by chromosome number, to create 20 somatic chromosome and one X chromosome file for each platform. Per chromosome files were converted to `.fasta` format.

Probe sequences from each platform were treated as analogous to short-reads from RNASeq data and aligned to each chromosome of the rat genome (Ensembl Rat genome assembly, 'Rnor_5.0', release 75, March 2012) (Consortium 2004) using a short-read, un-spliced alignment algorithm 'Bowtie' (Langmead, Trapnell et al. 2009) as implemented in the `Quasr` *R* package (Lerch, *et al* 2012). General methodology was adapted from Kim, *et al* (2011) (Kim, Patel et al. 2011). The unmasked genomic DNA sequences from Ensembl were used, where interspersed repeats and low complexity regions are retained, as this improved probe hits.

Aligned reads, `.bam` files, were handled using the `GenomicRanges` package (Lawrence, Huber et al. 2013) and visualised using the `Gviz` package (Hahn, *et al*, 2014). Probe overlaps were defined as probes sequences from Affymetrix (25-mer) or Illumina (50-mer) that were: a) contained entirely *within* or, b) had *any* sequence overlap with Agilent (60-mer) probes. Probes were only used if the same gene was mapped uniquely for each platform.

5.2.5: Human Affymetrix data set collection

The collection of Affymetrix microarray expression datasets profiling human cartilage and tendon samples followed the same protocol as outline above (**5.2.1-5.2.2**). The datasets retained for cross-species analysis are presented in **Appendix Table SD5.2**. Following pre-processing, re-annotation and cross-platform normalisation a human meta-set of 166 samples and 12,215 genes was prepared. Samples were obtained from the following platforms: Affymetrix Human Genome U133 Plus 2.0, U133A, Gene 1.0 ST and Gene 1.1 ST. A non-musculoskeletal tissue meta-matrix was also prepared from human liver microarray studies performed on the Affymetrix Human Genome U133 Plus 2 platform (n=150)

from 5 studies, **Appendix Table SD5.3**, profiling 19,851 transcripts. Samples were handled as described above.

5.2.6: Network comparisons of Affymetrix microarrays across species through the application of weighted gene co-expression network analysis (WGCNA)

Using a weighted gene co-expression network analysis it is possible to collate multiple data-sets into a single correlation matrix to define consensus network structures and modules which are strongly conserved or divergent across conditions, or species, and functionally annotate these modules. The protocol employed follows those defined in **Chapter 4**. The reader is referred to **Figure 4.2c** for a schematic overview of the methodology.

To establish universal gene identifiers and facilitate comparison across species rat gene identifiers were re-annotated with human Ensembl gene orthologs; the human meta-matrix was also re-annotated with Ensembl identifiers – only a single probe set per gene was permitted. After re-annotation this left rat and human meta-matrices of 10,221 and 11,904 genes respectively. Only identifiers that were common to both meta-matrices were retained and genes with a global variance less than 0.4 were removed to reduce noise and computational demands. This left 5278 genes for co-expression analysis. The general approach employed to develop cross-species co-expression modules is described by Miller, *et al* (2010) (Miller, Horvath et al. 2010). Consensus network and module generation was performed in `WGCNA` with the following changes to the default settings for consensus network generation: β=7, deepSplit=1, cutHeight =0.25 and a minimum module size of 30 genes. All other settings were left at default.

Sub-setting data sets to challenge network structures

To assess whether network structures would be robust to changes in data set each species meta-matrix was split into two groups consisting of a randomly selected sub-set of samples. For each group the average expression rank for each gene in the network was calculated. Within each species these ranks were correlated between the two random groups; ten iterations of random sample selection were performed. This was performed both using the 5,278 genes prepared for co-expression analysis. Two of the randomized sub-sets for each species were then used as inputs for full co-expression analysis to assess how robust the analysis was to sub-setting and changes in the input data sets.

Correlation of phenotypic traits to module eigengenes

Defining the relationship between phenotypic traits from the constituent samples within the individual data sets was undertaken as described in **Chapter 4, 4.2.1.** Binary classification tables were prepared for either the rat or human datasets as before. For the human data multi-dimensional scaling plots were prepared to define groups that were then used as phenotypic classifiers. For consensus module-trait relationships the lowest absolute value was defined as the correlation for a consensus module-trait pair if the two correlations had the same sign.

5.2.7: Class prediction analysis

Class prediction for culture subsets within the meta-sets was performed using the `pamr` package implemented in *R* and as described by Tibshirani, *et al* (2002), (Tibshirani, Hastie et al. 2002), **Figure 5.1d**. Briefly, this method employs a 'nearest shrunken centroids' approach to determine cohorts of genes that best characterise the defined classes. Modules defined by co-expression network analysis were used to filter the features (genes) for selection. Expression data

from rat or human were split into a randomized test group (10% of all samples for a meta-set) and a training set; the latter was used to develop the minimal gene signature. The gene signature for each species was cross-validated by testing against both species test groups. Principal component analysis was used to present separation of samples using the minimal gene signature.

All graphics, network visualization, functional annotation tools, and methodologies are consistent with previous descriptions.

5.3: Results

5.3.1: Cross-platform normalisation techniques

Multiple cross-platform normalisation strategies are available within R packages. To explore the impact on diverse samples alternative cross-platform normalisation strategies were applied to simulated data (not shown), 'seed' data (derived from real expression profiles) or real data to provide an objective analysis. Data sets derived from 'seed' expression values from the real data sets presented similar findings using either DWD or XPN normalisation methods with global increases in the average Pearson correlation between all samples and platforms, **Figure 5.2**.

For real data sets BMC and Genenorm (z-score normalisation) presented with identical distributions where the majority of correlations between samples were centered on zero, i.e. not correlated. In contrast, the DWD method again produced a high correlation, >0.98, across all samples and platforms, but with Affymetrix Genome 230 and GeneST 1.0 arrays showing the greatest overlap of correlation distributions, **Figure 5.3**. Using the real data sets (n = 12) it was not possible to apply the XPN method. Principal component analysis of real data, **Figure 5.4**, found that BMC and DWD methods resulted in a strong clustering of two platforms, Affymetrix 230 and Agilent, regardless of whether the samples were considered 'case' or 'control'. This was not replicated for z-score normalised data where 'case' and 'control' samples were still separated between Affymetrix and Agilent platforms.

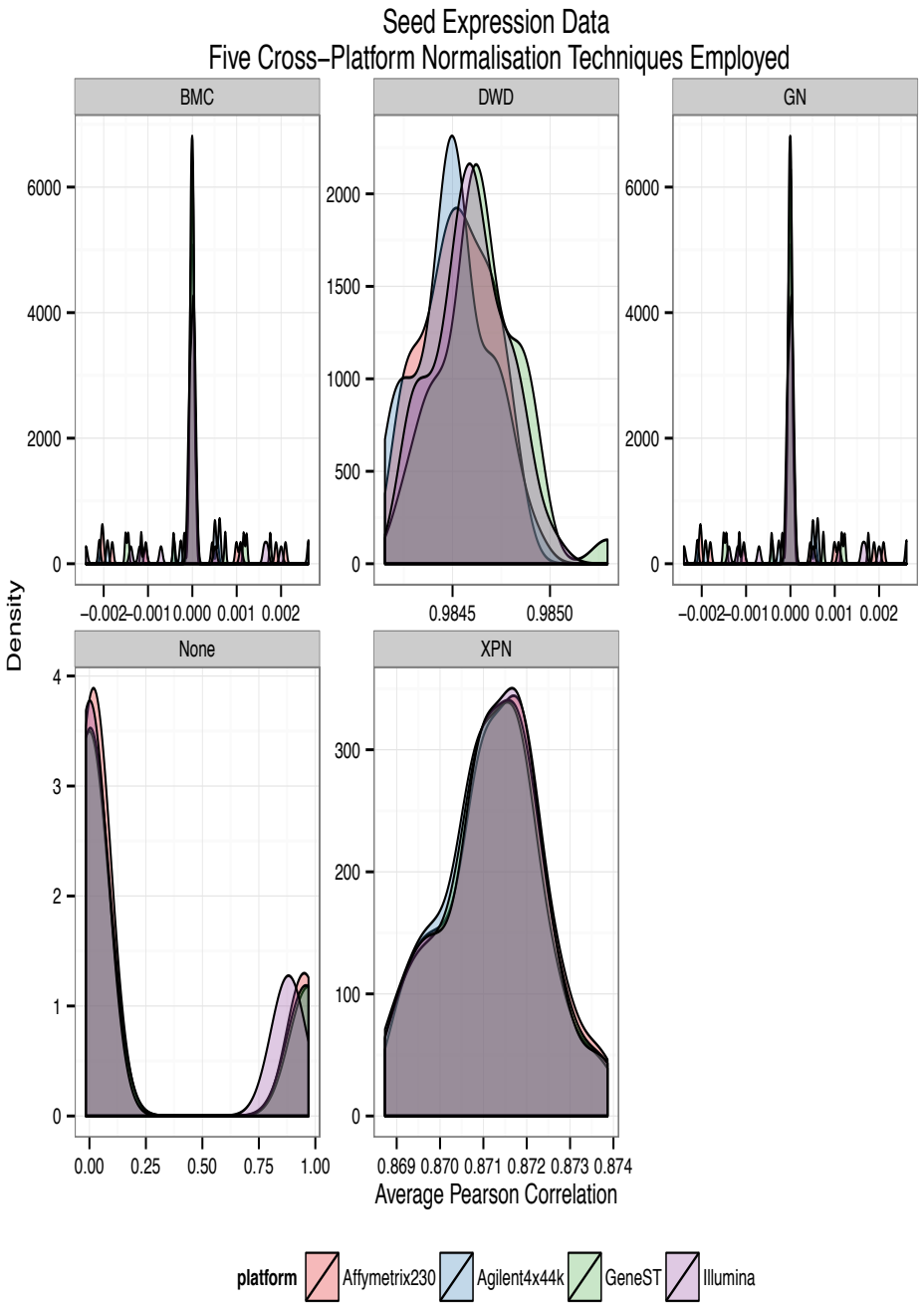


Figure 5.2 – Correlation density plots: Five cross-platform normalisation strategies were considered using artificial data sets created using real data-sets as a 'seed' around which the distribution of expression values was created. Normalisation algorithms of differing complexity were used: BMC; DWD; Genenorm; None; or XPN. Data from four platforms were considered, representative of those available in public repositories – see **figure legend** (platform). Normalisation algorithms such as XPN and DWD have assumptions that the data to be normalised arises from the same sample type, but analysis undertaken on different platforms. As such, these algorithms result in global correlations of >0.8 across samples known to come from different conditions. Genenorm and BMC use a standardisation method where the expression data is 'studentised', i.e. mean=0, SD+/- 1.

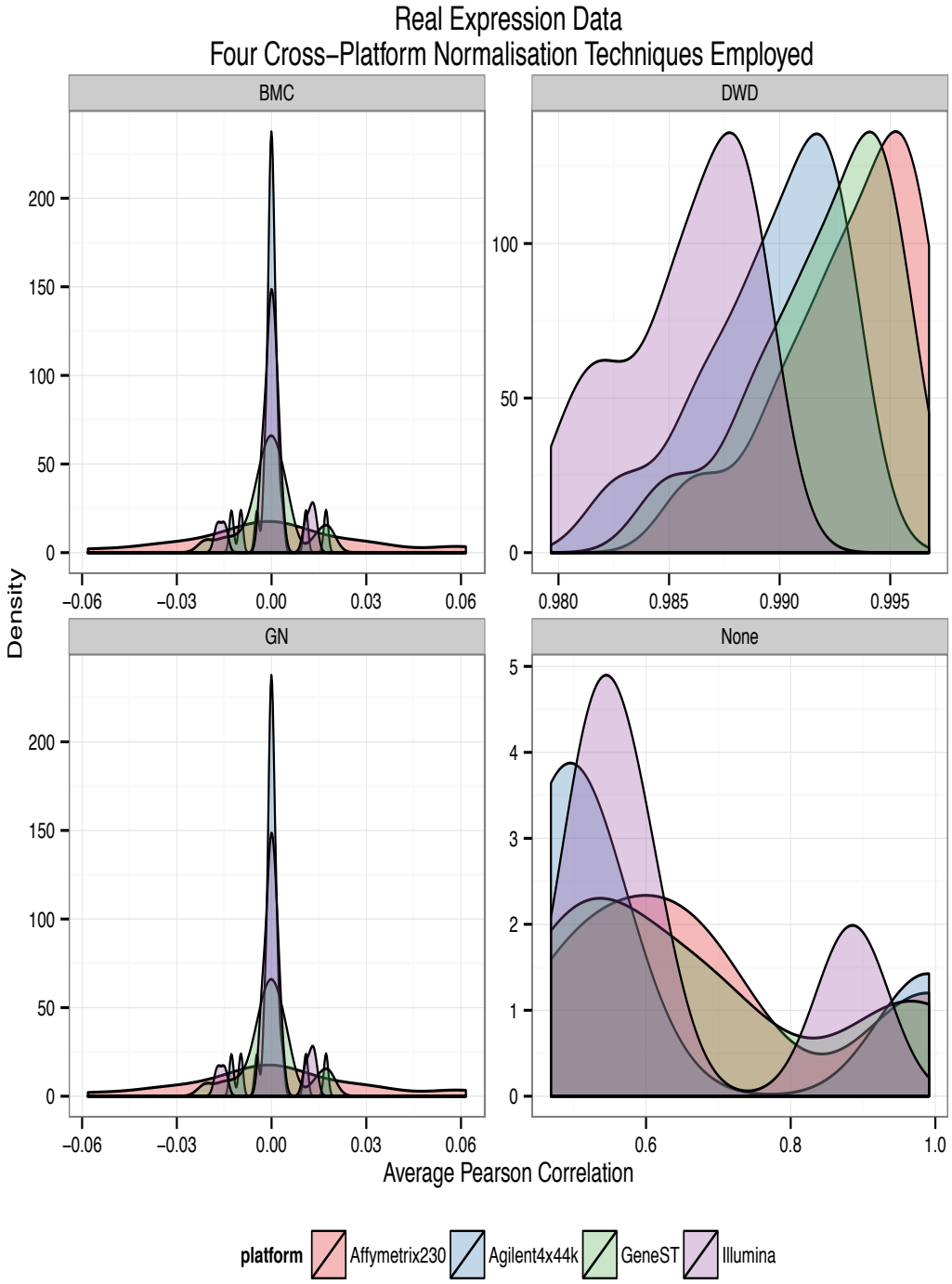


Figure 5.3 – Correlation density plots: Four cross-platform normalisation strategies were considered using **real** data sets creating consisting of three case and three control studies. Normalisation algorithms of differing complexity were used: BMC; DWD; Genenorm; None. Data from four platforms were considered, representative of those available in public repositories – see **figure legend** (platform).

Application of the DWD algorithm to all platforms resulted in high cross-platform correlations, >0.98, with the two Affymetrix platforms (230 and GeneST) showing the highest correlation. Applying no normalisation technique resulted in variable correlations including some very high correlations approaching cor=1.

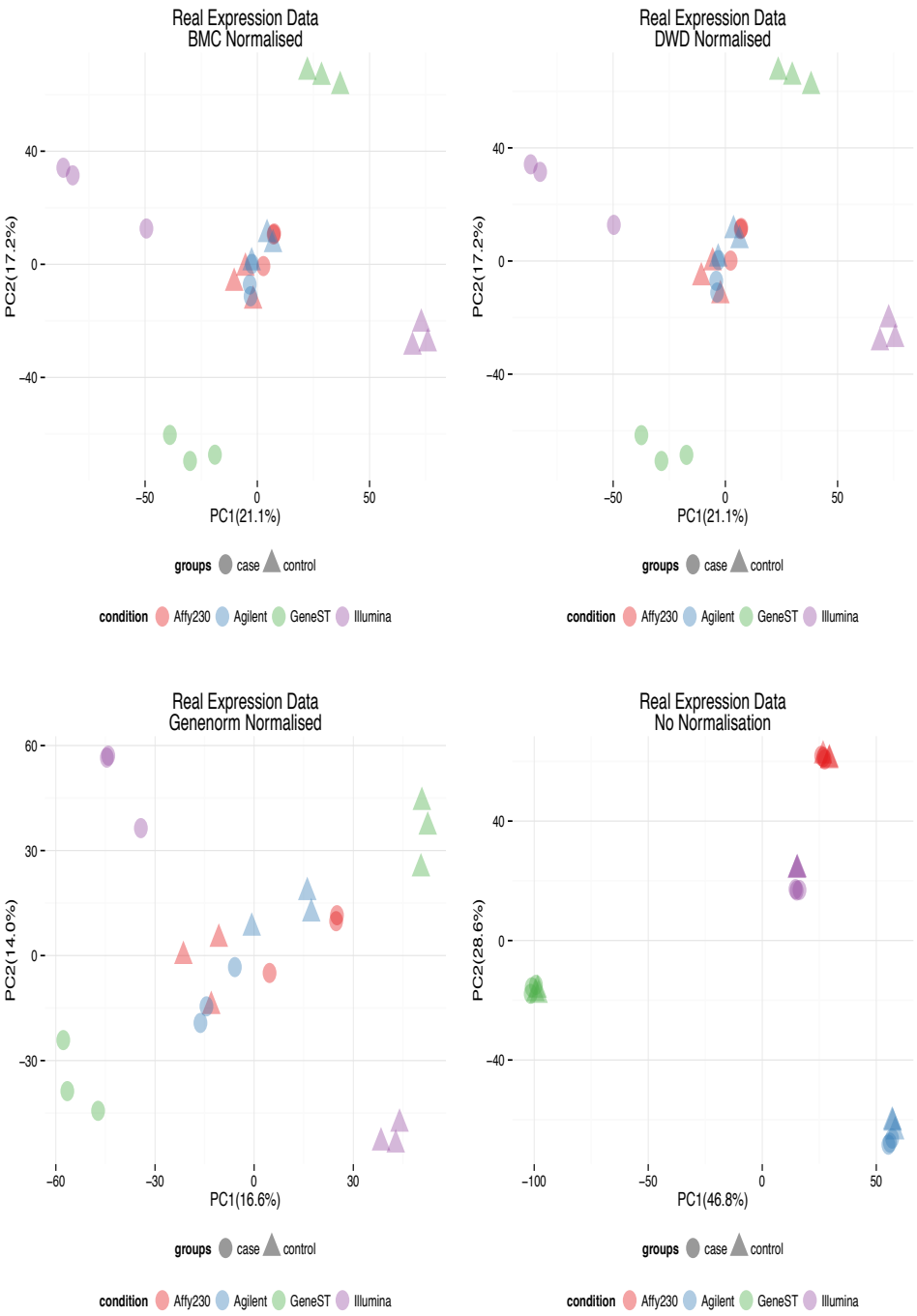


Figure 5.4 – Principal Component Analysis – Data arising from **real** microarray expression data sets derived from cartilage samples applied across four platforms. Four cross-platform normalisation techniques were considered as per **Figure 5.3**. PCA explained the greatest variance in the first two components for BMC and DWD strategies. These methods strongly clustered data from Affymetrix Genome 230 and Agilent 4x44k arrays, including both real case and control data. Genenorm normalised data clustered in the same manner as for BMC and DWD, but with a smaller total PCA percentage, but retained less stringent clustering of the case and control arrays across the two arrays. Combining the expression data without applying a cross-platform normalisation strategy resulted in strong clustering of arrays by platform only.

To investigate whether cross-platform normalisation was required between platform generations from a single manufacturer the steps outlined above were applied to three generations of Affymetrix platform: a) GeneChip Rat 1.0 ST; b) GeneChip Rat 2.0 ST; c) Rat Genome 230 2.0 arrays. As before, BMC and Genenorm resulted in a similar distribution of correlation values across the expression data; DWD resulted in a global cross-platform increase in correlation, **Figure 5.5**. Notably, not employing a cross-platform normalisation technique for Affymetrix platforms resulted in a higher cross-platform correlation than compared to the previous multi-platform analysis.

By principal component analysis DWD and BMC normalisation clustered all samples from the Affymetrix 230 and Affymetrix ST 1.0 arrays, **Figure 5.6**. This was considered over-smoothing of the data and resulted in groups that were not biologically plausible with respect to the sample descriptors from the original studies. *Z*-score normalisation resulted in less-stringent clustering of the replicates and, although less of the data variation was described by the first two components, this was considered to be more biological plausible clustering of samples, **Figure 5.7**.

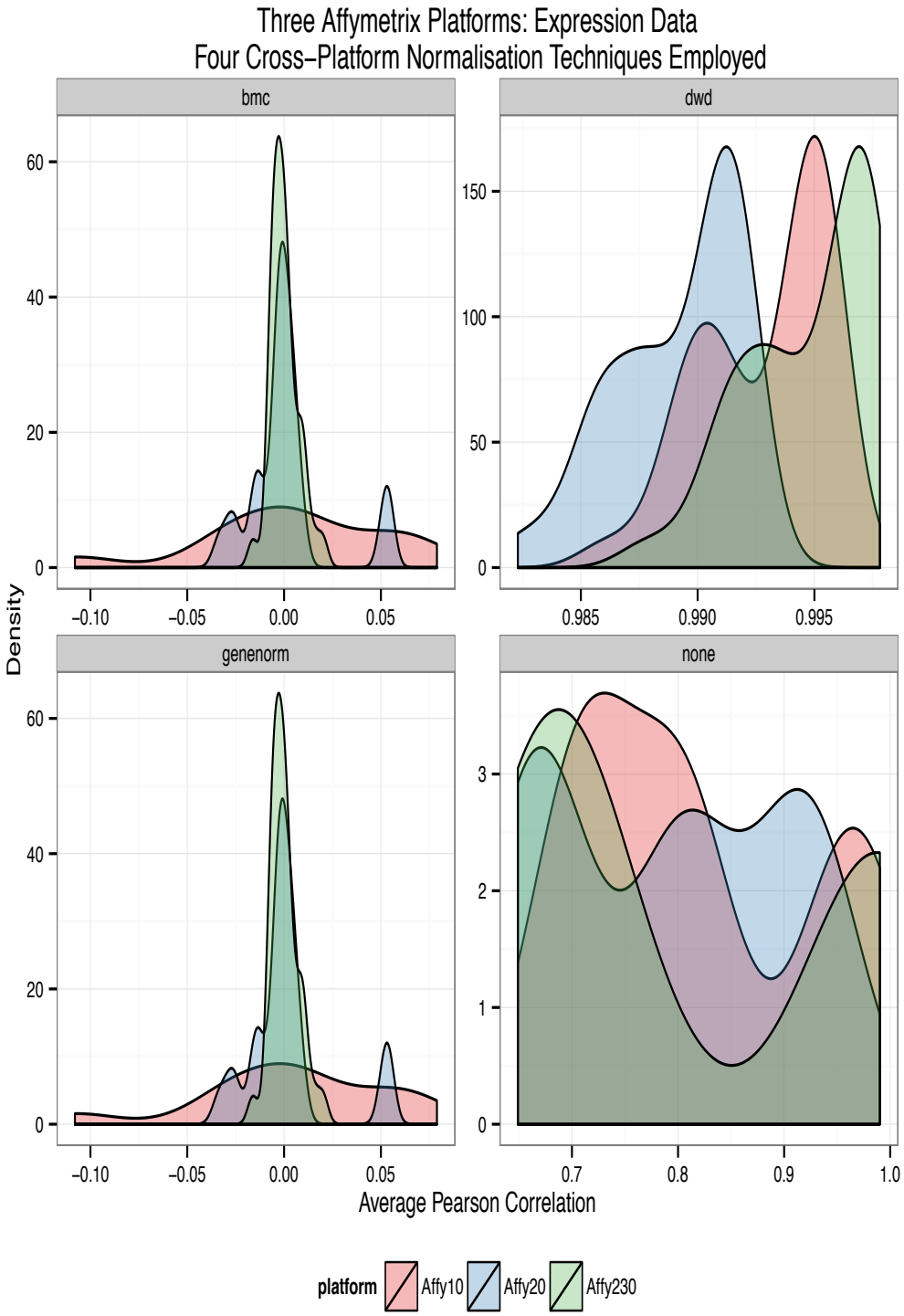


Figure 5.5 – Correlation Density Plots– Application of normalisation strategies presented in **Figure 5.3** to arrays derived only from Affymetrix platforms – a) GeneChip Rat 1.0 ST; b) GeneChip Rat 2.0 ST; c) Rat Genome 230 2.0 arrays – see **figure legend** (platform). In general the findings were comparable with findings across multiple platforms. DWD resulted in high inter-array correlations. No normalisation resulted in moderate to high correlations between platforms, but these values were more widely distributed. BMC and Genenorm strategies resulted in distributions comparable with previous findings.

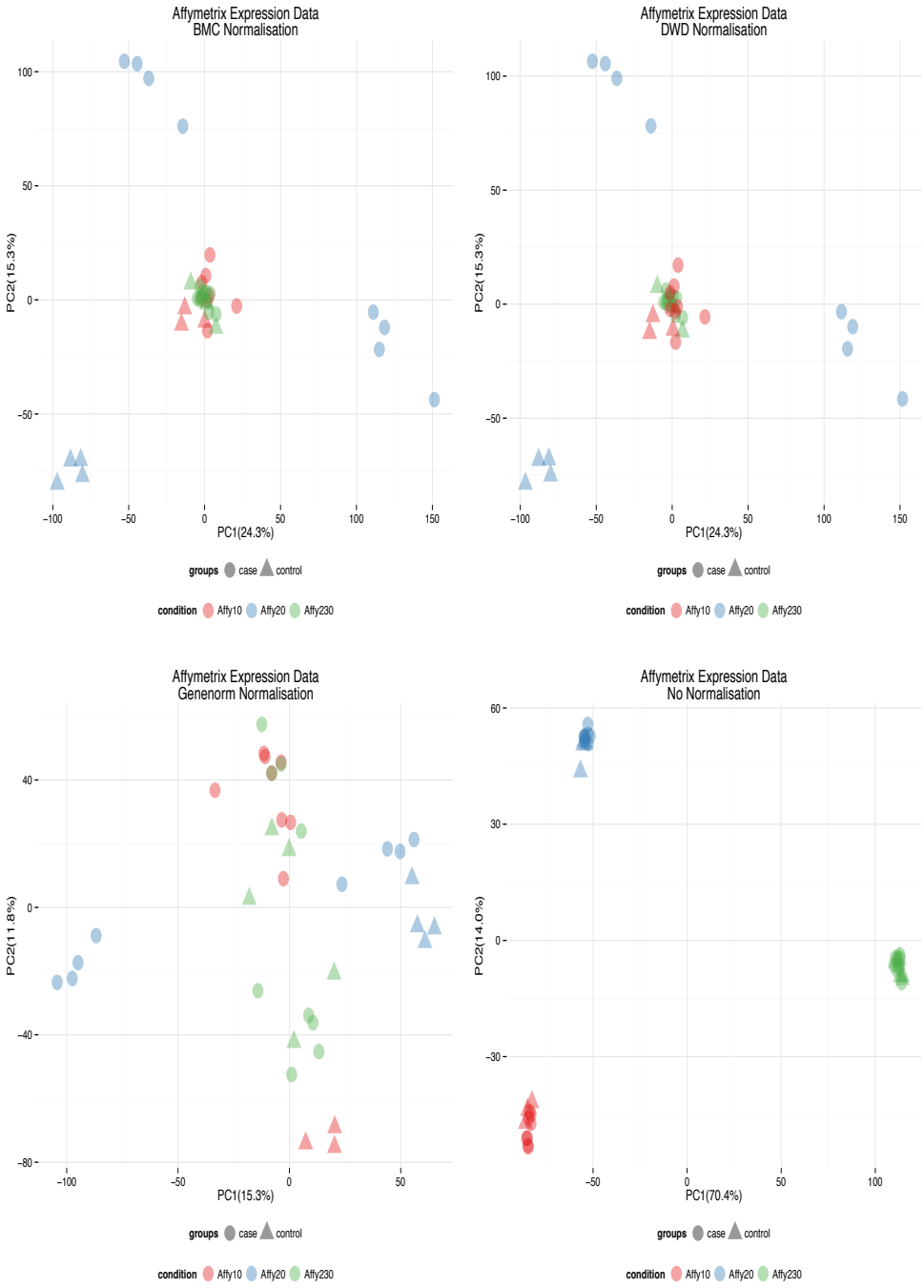


Figure 5.6 – Principal Component Analysis - as demonstrated in **Figure 5.3** the DWD and BMC strategies resulted in strong clustering of all data from Affymetrix 1.0 ST and Genome 230. In contrast, data derived from **Chapter 3** Affymetrix 2.0 ST arrays were divergent from this central cluster. No normalisation strategy resulted in clustering by Affymetrix platform generation. Genenorm normlisation resulted in a broadly similar distribution of samples, but more relaxed clustering was evident with separation of case and control samples – see **figure legend** (groups, condition).

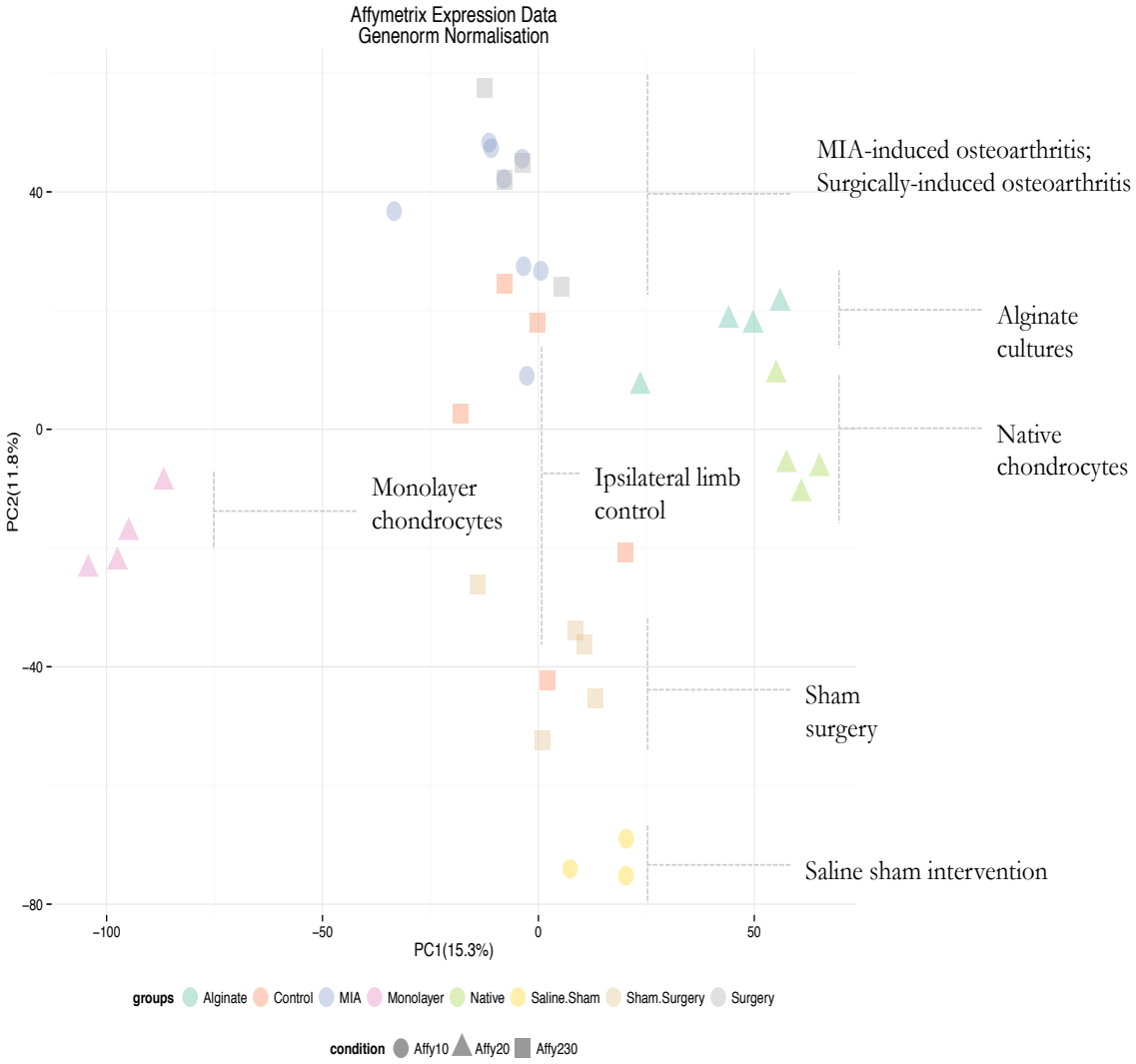


Figure 5.7: Annotated principal component analysis plot –

Genenorm normalised data from Affymetrix microarrays (**Figure 5.6**) annotated for sample origin – **figure legend** (groups, condition).

Unlike DWD and BMC normalisation strategies Genenorm normalisation resulted in separation of samples in a biologically relevant manner, i.e. consistent with sample descriptors from the source data sets.

Perturbed chondrocytes, either from inflammatory or *in vitro* conditions, clustered separate from sham interventions, control cartilage and matrix-depleted native chondrocytes.

5.3.2 Sequence- and target-matching of probes across platforms curtails numbers of available probes for data integration

Given the dearth of publically available expression profiles for cartilage and tendon including data from multiple platforms would be useful. To investigate whether sequence- and target-matching probes from different platform manufacturers would be a valid methodology for improving reproducibility of the analysis oligo probe sequences from two Affymetrix arrays (1.0 GeneST and Genome 230), one Illumina, and one Agilent platform were aligned to the *Rattus norvegicus* genome and probes were considered sequence- and target-matched across platforms if the shorter probes from the Affymetrix (25-mer) or Illumina (50-mer) platforms could be unambiguously matched *within* the larger Agilent probes (60-mer) uniquely aligned to the genome. The number of sequence matches, or 'hits', within Agilent probes was highest across all chromosomes with the shorter 25-mer Genome 230 probes. Comparatively, GeneST 1.0 probes, of the same length, had few hits within Agilent probes in most chromosomes; in some cases there were no hits in a chromosome, **Figure 5.8**

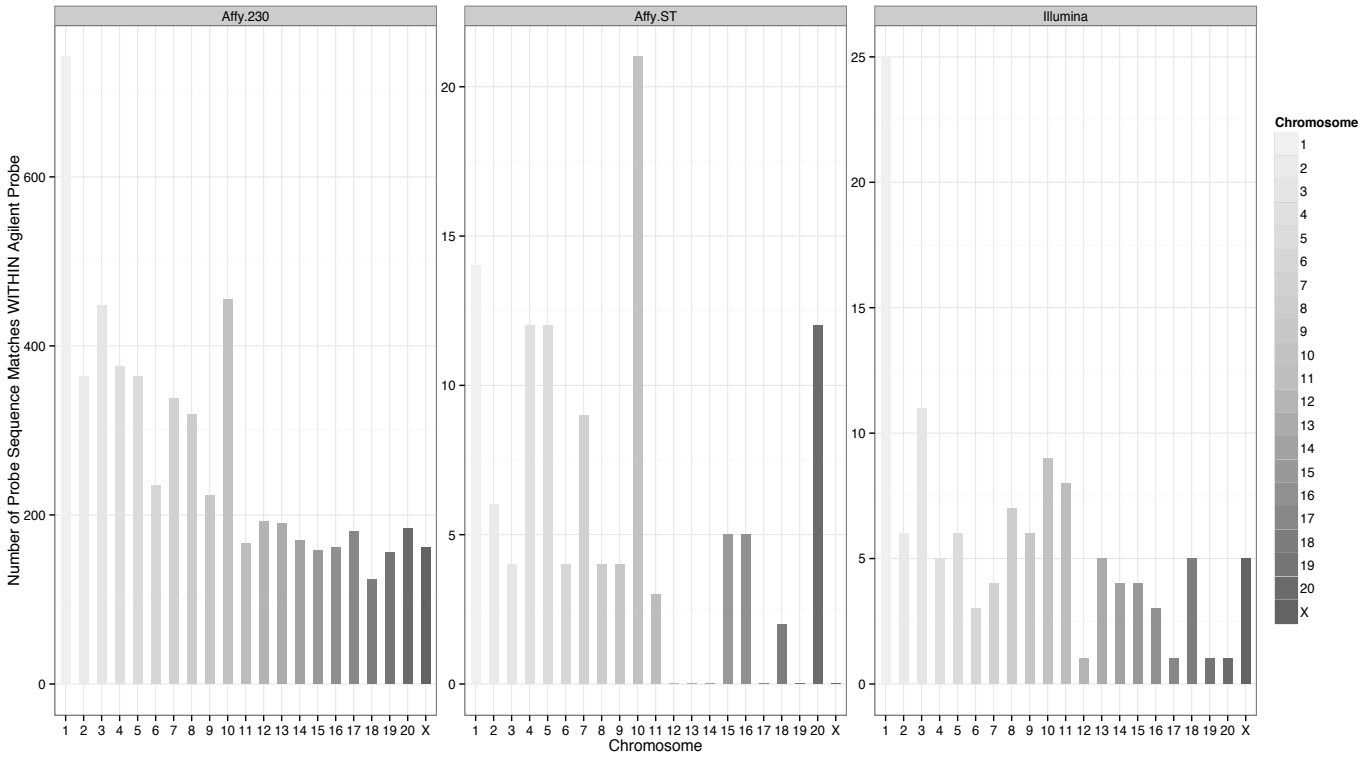
Given this very stringent definition of sequence matching severely limited the number of matching probes a more relaxed definition was considered, which allowed *any* overlap with an Agilent probe to be considered a hit. Generally, the number of matched probes increased considerably for Affymetrix 230 and Illumina probes, but had little impact on the GeneST 1.0 probes, **Figure 5.9**.

Multiple Affymetrix probes (probe-sets) are annotated to the same gene; mapped hits for all probes were re-annotated to define uniquely represented genes that were matched across all platforms. Given the poor number of hits for the GeneST 1.0 array the consensus genes for only Affymetrix 230, Illumina and

Agilent were considered with duplicates removed. After annotating with Entrez gene identifiers the number of unique identifiers across platforms was associated with 620 (Agilent), 615 (Illumina) and 616 (Affymetrix 230) unique genes matched.

Sequence matching of probes across three platforms resulted in considerable loss of gene expression profiling capability; given that just over six-hundred probes could be uniquely identified as confidently hybridizing to the same region of a transcript this methodology was considered as an unacceptable way of tackling a cross-platform meta-analysis using multiple platforms.

The simplest normalisation strategy, z-score normlisation, did not over-smooth the data, resulted in clustering that was less stringent and described less of the variability in the data, but was more relevant to the biological understanding of the samples. Consequently, data merging was undertaken entirely with microarrays developed by Affymetrix using 25-mer oligo probes.



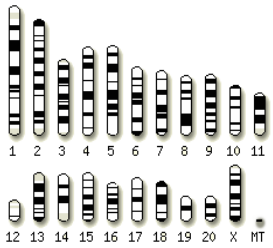


Figure 5.8 – Histogram of probes hits **within** Agilent probes per chromosome - Affymetrix platforms use 25-mer oligo probes, but have multiple probes for each gene or exon depending on the platform. There is a notable difference between the number of Affymetrix 230 probes that lie within a 60-mer Aglient probe (and so verified as sequence-matched) and the Affymetrix GeneST 1.0 array where, in some chromosomes, there are no probes that lie within an Agilent probe. Illumina arrays, like Agilent, have only one probe per gene and are 50-mer oligos – few of these probes lie completely within an Agilent probe on the same chromosome. The number of overlapping probes do not necessarily following the approximate size of each chromosome (**inset**).

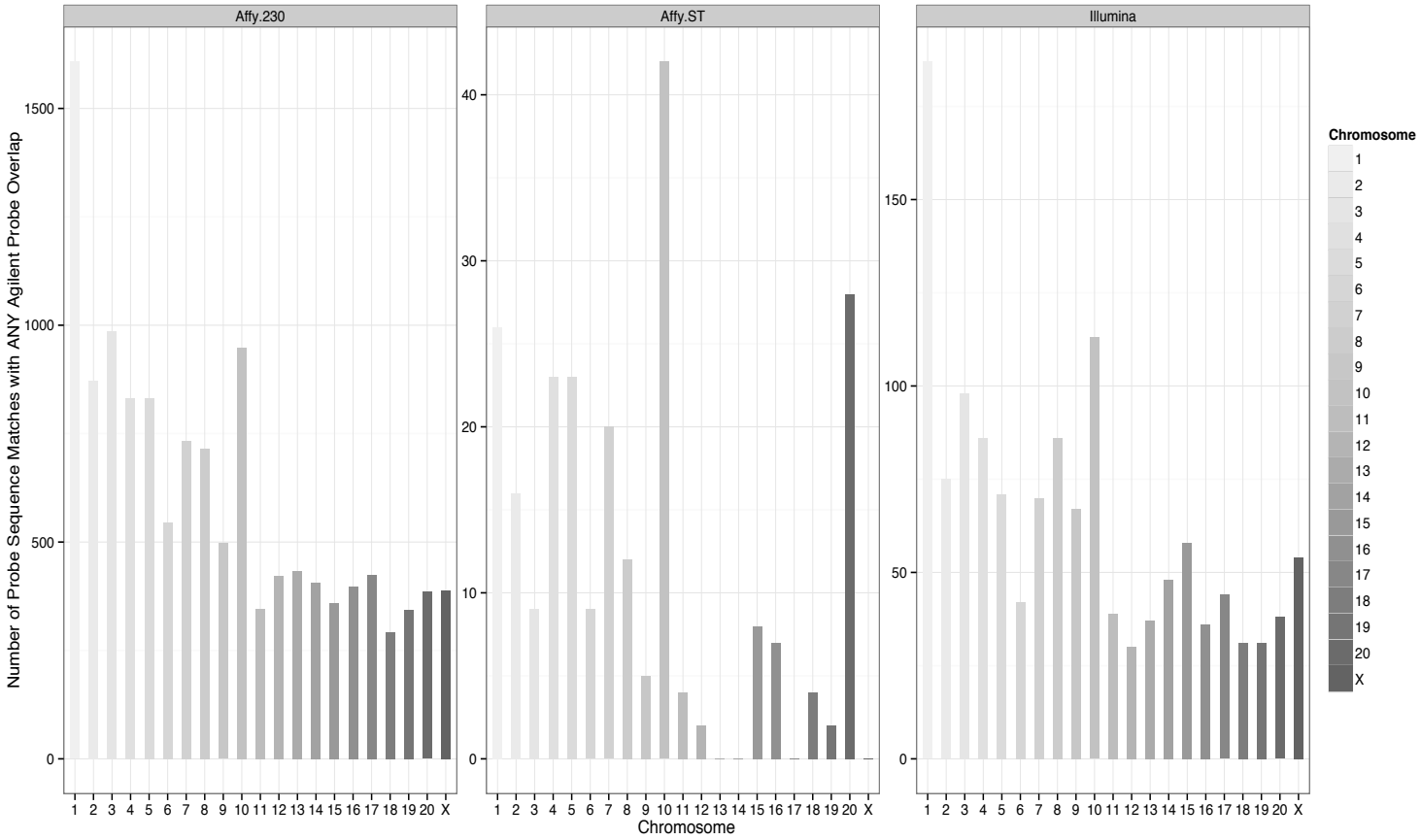


Figure 5.9 – Histogram of probes overlapping Agilent probes per chromosome – relaxing the qualification for sequence matching to include ANY overlap between Agilent probes and those from other platforms increased the number of probe hits (y-axis) across sequences chromosomes (x-axis). It was still the case in some chromosomes that there was no overlap called between Agilent probes and Affymetrix Gene ST 1.0 arrays

5.3.3: Analysis of co-expression networks across rat and human data identifies cross-species preservation of functional gene modules

Constructing networks for rat and human meta-sets

To develop an insight into musculoskeletal disorders co-expression networks were developed to survey the general responses of chondrocytes and tenocytes to system perturbations. After filtering by defined inclusion qualifiers the analysis included 336 microarray samples consisting of cartilage and tendon studies from 166 human and 170 rat investigations. A single gene expression meta-matrix was prepared for rat or human data. Each meta-matrix was assessed to determine whether it met an approximately scale-free topology, **Figure 5.10**. Strong scale-free topology was evident. Initially species-specific single networks were prepared and, using the topological overlap measurement described in **Chapter 4**, hierarchical clustering was used to group genes with high co-expression relationships into modules. After conversion of rat gene identifiers to converted to human orthologs, intersection of common genes and non-specific filtering to facilitate computational analysis a set of 5278 genes common to both species meta-matrices remained for co-expression analysis. All further analysis uses this gene set.

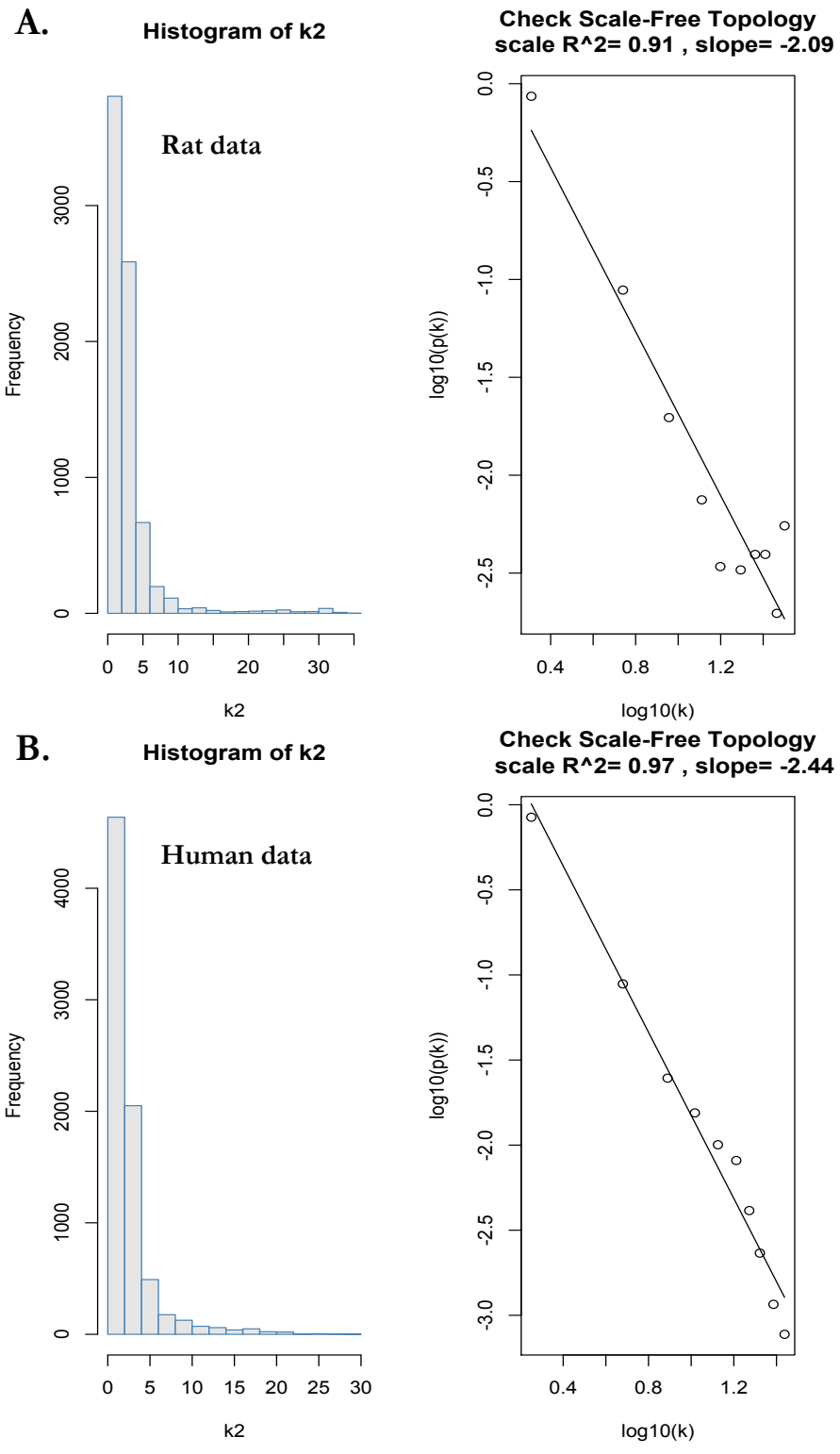


Figure 5.10: Visual assessment of scale free topology after non-specific filtering for rat (A) and human (B)– Histograms: Frequency (y-axis) of gene connectivities (x-axis). The majority of genes have few connections; Log-log plot of the connectivity values shown in the associated histogram with fitted linear model. The R^2 value is considered an index of the scale-free network topology. For the rat (7616 genes) and human (7759 genes) meta-sets there was good evidence of scale-free network topology.

5.3.4 Assessment of general network properties

Networks and module definitions are robust to the choice of data-set

As the constituent data sets for each species meta-matrix came from disparate sources it was necessary to demonstrate that networks would be robust to changes in the data sets. The single species meta-matrices were randomly split to contain half the number of total samples (rat, 85; human, 83) and were processed using the same protocol as described previously. Ten iterations of random data sets were prepared and analysed.

Correlation of the mean expression rank between random data sets in the rat were, in 9/10 iterations, positively correlated (range: R=0.04 - 0.36; p=1.3e-05 to 2.6e-191) and for all iterations using human data (range: R=0.1 – 0.43; p=3.7e-02 to 1e-200). In the majority of random permutations of the gene expression data in either rat meta-matrix the data was internally comparable with reference to published guidelines for WGCNA.

Gene expression rank was also correlated between species. Correlation was positive for rat versus human expression data (R=0.13, p=2.5e-21), but negative for the human cartilage/tendon meta-matrix versus the human liver meta-matrix (R= -0.091, p=3.5e-11) indicating cartilage/tendon meta-matrices across species were more comparable than between tissues within the same species.

Two random data-set pairs from each species were chosen, consensus networks prepared on these sub-sets and module preservation considered. Module eigengene differential analysis, module functional annotation and consensus hub gene definitions were consistent with the findings defined in the succeeding sections (data not shown).

These findings demonstrate that within each species meta-matrix there was no significant data bias; networks derived from randomized sub-sets of the meta-matrices resulted in highly comparable module structures and definitions, functional annotation and hub gene prediction. It was concluded that the findings did not arise as a function of the meta-matrix construction alone.

Functional annotation of rat network modules

The rat network consisted of eleven modules, and one additional module containing unassigned genes (grey). These modules are described in terms of their functional annotation in **Table 5.2**. Module colours are not interchangeable between rat and human. In general terms these functional annotations were very similar to those previously identified for cartilage and tendon samples, **Chapter 4**.

Human modules share rat functional annotation

Network generation for the human meta-matrix defined six distinct modules, plus one module containing unassigned genes. Functional annotation of modules, **Table 5.3,** demonstrated significant enrichment for terms also found in rat modules. For example, the H3 module was described by terms relating to: 'immune system process' and 'defense response'. The H2 module was described by terms relating to 'muscle process', 'skeletal muscle contraction' – terms previously highly correlated with native tendon. The H5 module was strongly associated with terms relating to 'cytokine activity' and 'chemokine signalling' pathways.

Table 5.2: Rat meta-matrix specific modules. Genes assigned to modules in **Figure 5.13** were assessed for functional annotations using an over-representation analysis in DAVID. Terms that are not significant are indicated (ns) and are provided for completeness. The most significant term (Bonferroni-corrected) for each of the GO families is provided alongside over-represented KEGG pathways. NA – not available

	Biological Process	Metabolic Function	Cell Compartment	KEGG Pathway
R1	Cell cycle (8.2e-44)	DNA binding (3.4e-6)	Nuclear part (1.3e-34)	DNA replication (8.4e-27)
R2	Anatomical structure development (1.0e-6)	Actin binding (1.7e-5)	Extracellular matrix (5.7e-9)	NA
R3	Cellular protein metabolic process (ns)	Transferase activity, glycosyl groups (2.0e-3)	Endoplasmic reticulum (9.7e-9)	NA
R4	Ossification (ns)	Growth factor activity (ns)	Organelle (ns)	Hedgehog signaling pathway (7.2e-3)
R5	Heme biosynthetic process (1.3e-2)	Lyase activity (ns)	Cortical cytoskeleton (ns)	Porphyrin and chlorophyll metabolism (6.7e-3)
R6	Response to external stimulus (7.2e-17)	Cytokine activity (2.9e-11)	Extracellular space (5.8e-16)	Cytokine-cytokine receptor interaction (1.2e-9)
R7	Immune response (1.8e-9)	Peptidase activity (ns)	TAP complex (4.1e-3)	Toll-like receptor signalling pathway (ns)
R8	Immune system process (3.3e-23)	Molecular transducer activity (5.4e-6)	Plasma membrane (1.2e-16)	Lysosome (4.9e-3)
R9	Signal transduction (3.7e-9)	Molecular transducer activity (4.7e-8)	Plasma membrane (3.9e-5)	Cytokine-cytokine receptor interaction (3.1e-2)
R10	Generation of precursor metabolites and energy (2.5e-51)	NADH dehydrogenase activity (1.1e-16)	Mitochondrial part (5.6e-54)	Parkinson's Disease (2.4e-24)
R11	Skeletal system development (5.4e-7)	Extracellular matrix structural component (3.3e-3)	Extracellular region part (5.9e-8)	NA

		Biological Process	Metabolic Function	Cell Compartment	KEGG Pathway
H1		Cell cycle (4.6e-48)	DNA binding (2.3e-3)	Nuclear part (1.5e-28)	DNA replication (1.5e-22)
H2		Muscle contraction (2.0e-23)	Cytoskeletal protein binding (5.9e-5)	Actin cytoskeleton (8.6e-10)	Insulin signaling pathway (4e-2)
H3		Immune system process (4.7e-19)	Molecular transducer activity (4.4e-4)	Lysosome (1.9e-6)	Antigen processing and presentation (2.7e-2)
H4		Blood vessel development (1e-9)	Calcium ion binding (ns)	Plasma membrane (8.4e-8)	Notch signaling pathway (ns)
H5		Response to external stimulus (6.6e-12)	Cytokine activity (7.3e-12)	Extracellular space (5.1e-8)	Cytokine-cytokine receptor interaction (7.3e-12)
H6		Interphase of mitotic cell cycle (ns)	Oxidoreductase activity (ns)	Cell fraction (ns)	Cell Cycle (ns)

Table 5.3: Gene ontology descriptors of the human meta-set modules defined in **Figure 5.14**. For each of the eigengene modules for the human cartilage and tendon meta-set the most significant biological process, metabolic function and cellular compartment gene ontology annotations are provided. The Bonferroni-corrected *p*-value is provided in parentheses; *ns* indicates $p > 0.05$ and annotations are provided for completeness. The 'red' module is the only human set-specific module, but shows no enrichment for gene ontology terms.

Generating consensus networks between rat and human meta-sets

In order to consider the preservation of network structure across species, beyond functional annotation, a consensus network was prepared. Across a range of soft-threshold values β=7 was considered suitable for both data set, **Figure 5.11**. Five consensus modules were defined, **Figure 5.12**.

Significant module overlap exists between rat and human network modules

The modules defined in single networks for the rat (**Figure 5.13**) and human (**Figure 5.14**) meta-matrices were assessed for overlap with the new consensus network modules. In general, modules from each species-specific network had a corresponding module in the consensus network and had significant enrichment for genes found in those modules using a permutation test to define a module preservation summary z-score, **Table 5.4**. In **Figure 5.15** the relative overlap between species-specific modules and consensus modules is summarized.

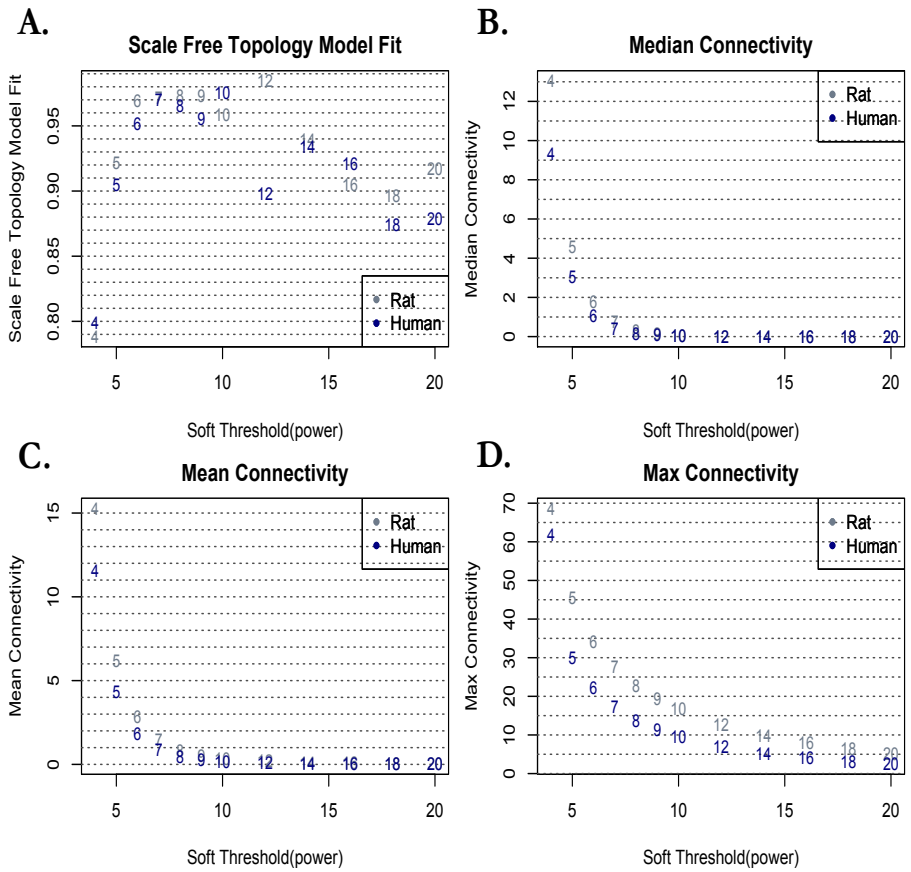
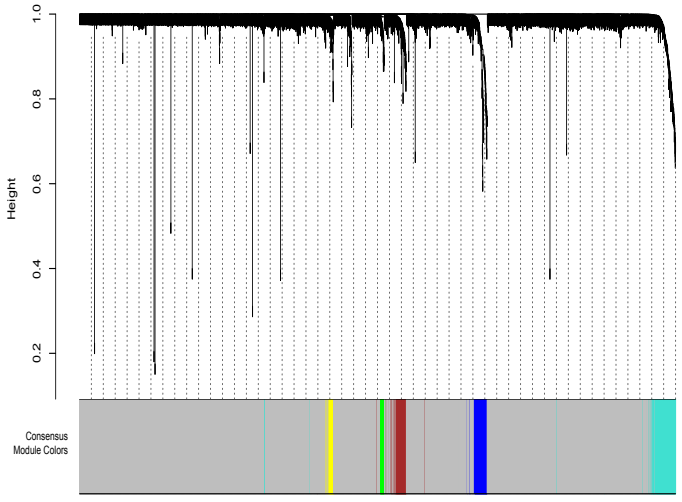


Figure 5.11 (upper panel): Determination of soft-thresholding powers for rat and human meta-sets– Plots show the summary network indices (y-axes) as a function of the soft-thresholding power (x-axis) for 5278 common genes. An approximate scale-free topology is reached at beta=7 for both sets (**A**). This was also chosen as it was the lowest power that still approximated the criterion for scale-free topology without losing the summary connectivity (**B-D**), which declines rapidly with increasing soft-thresholding power.

Figure 5.12 (lower panel): Hierarchical clustering gene dendrogram and consensus module definition: Rat and human meta-set analysis. Five consensus modules were defined, genes in each module are assigned the same colour (colour band below dendrogram). Grey areas represent unassigned genes.



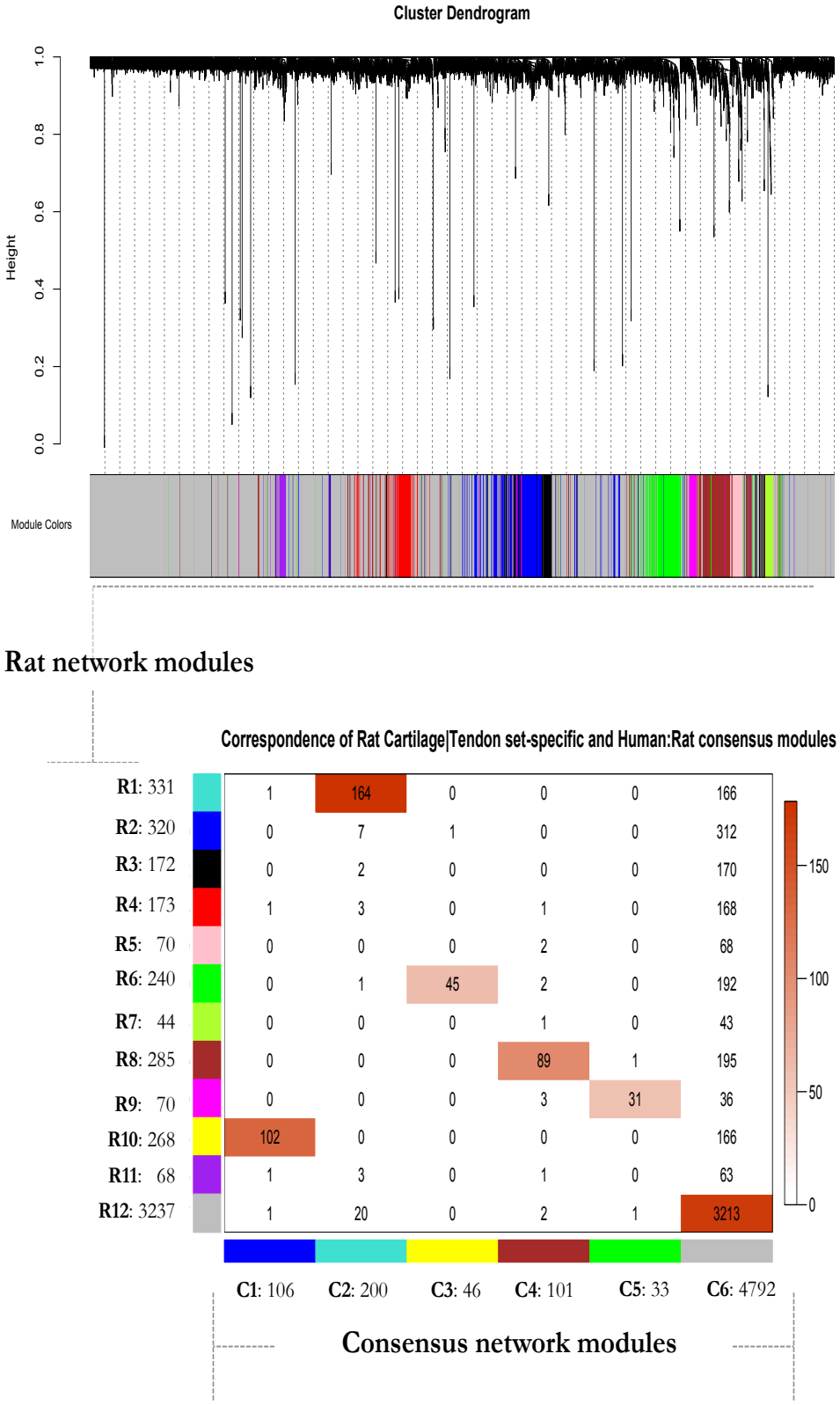


Figure 5.13: Rat meta-set modules (**top**) show multiple module eigengenes in the network. Relative to the rat–human consensus network several modules are not represented within the consensus network (**below**). Alternative alphanumeric annotations for rat and consensus modules are shown and these are used in the text. The total size of each module is indicated beside each module code. Degree of overlap between modules is represented in the graduated colour bar

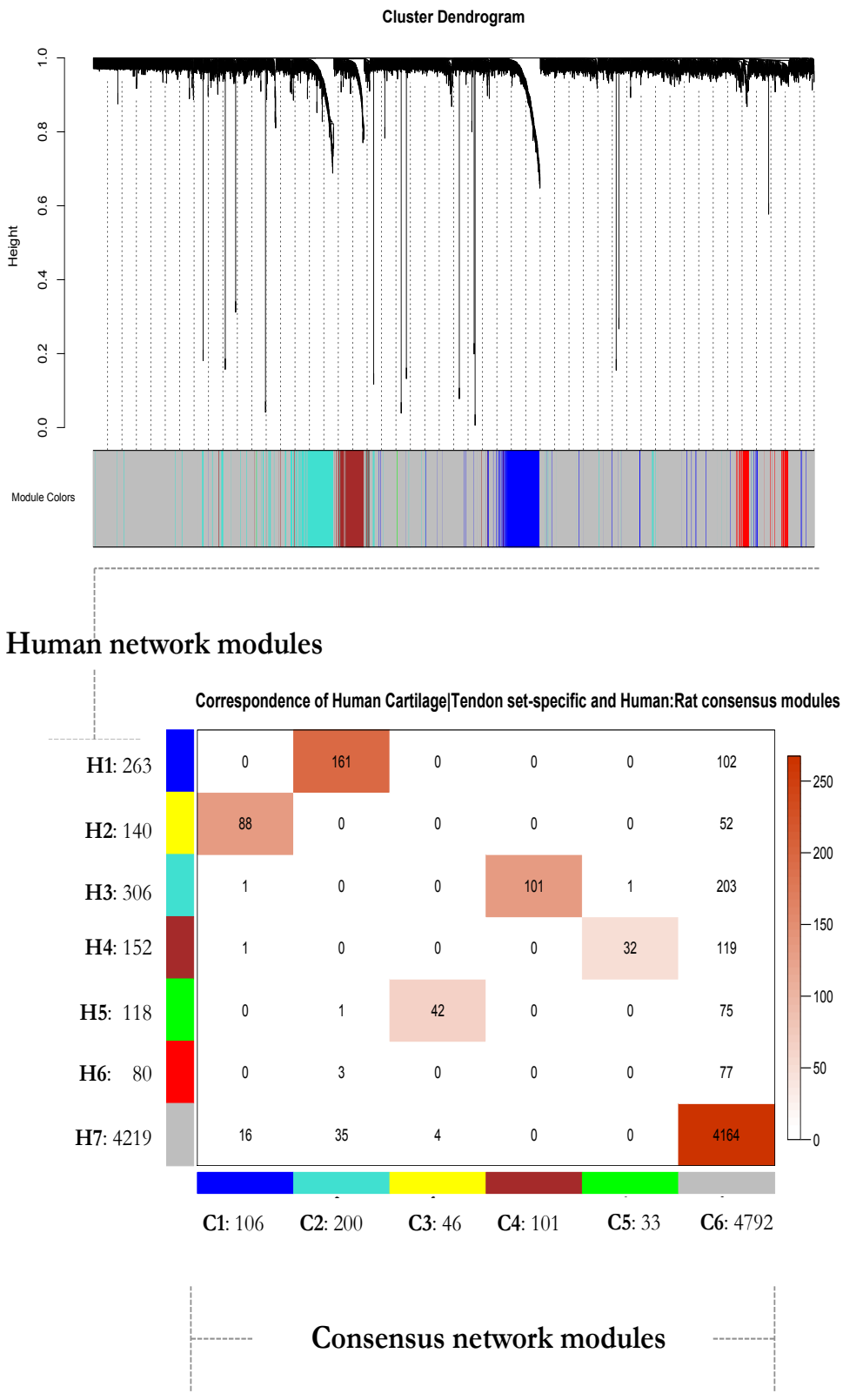
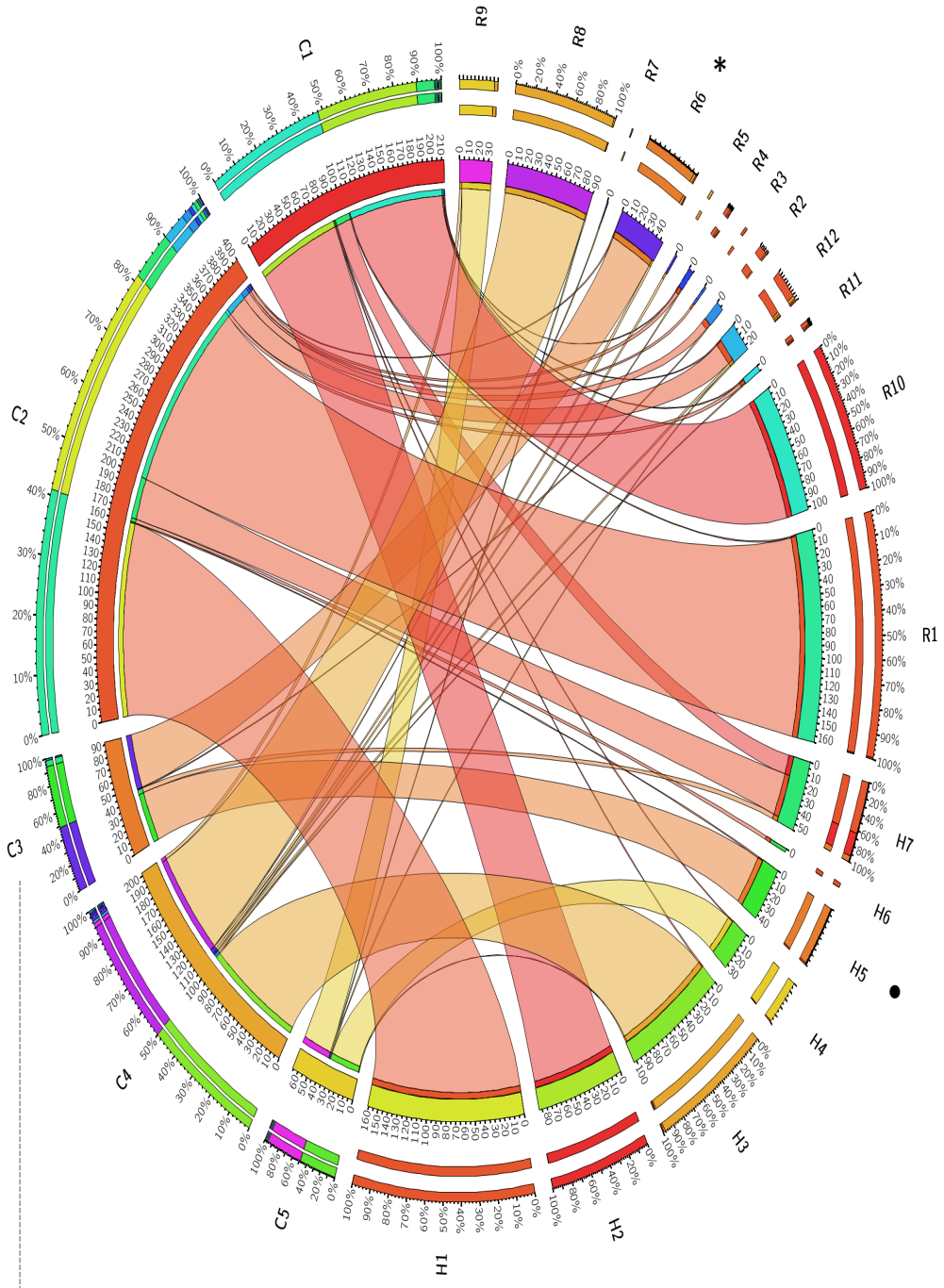


Figure 5.14: a) Human gene dendrogram and module definitions - b) Overlap of human module eigengenes with those identified in the consensus network – all modules from the human set-specific modules (rows) have an assigned module within the consensus network (columns), except the red module. Degree of overlap between modules is represented in the graduated colour bar.

Table 5.4: Module preservation table – Using the defined module overlaps in **Figure 5.13 and 5.14** the table indicates the module equivalents for both rat and human modules and the consensus modules. The module preservation score is based upon the rat module eigengenes. Where a module is species-specific this is indicated by the absence of an equivalent in the other module columns. The consensus gene ontology is provided to demonstrate the module equivalence across meta-sets. Where an annotation is not significant this is indicated (ns), otherwise annotations indicate the most significant terms for the consensus module genes.

Human Module	Rat Module	Consensus Module	Module Preservation Z-score (genes in rat modules, Bonferroni log10 p-value)	Consensus Gene Ontology BP | MF | CC | KEGG pathway (Bonferroni p-value <0.05)
Yellow (H2)	Yellow (R10)	Blue (C1)	29.6 (268, -271.2)	Muscle contraction | cytoskeletal protein binding | contractile fibre part | cardiac muscle contraction
Blue (H1)	Turquoise (R1)	Turquoise (C2)	36.7 (331, -422.9)	Cell cycle | chromosome | DNA binding | DNA replication
Green (H5)	**Green (R6)**	**Yellow (C3)**	**15.5 (240, -52.8)**	**Response to wounding | cytokine activity | extracellular space | NOD-like receptor signaling pathway**
Turquoise (H3)	Brown (R8)	Brown (C4)	24.1 (285, -163.9)	Immune system process | molecular transducer activity | plasma membrane | antigen presentation and processing
Brown (H4)	Magenta (R9)	Green (C5)	13.4 (70, -54.6)	Signal transduction | plasma membrane | signal transduction activity | neuroactive ligand-receptor interaction
-	Green-yellow (R7)	-	14.1 (44, -64.4)	Immune response | Peptidase activity ^{ns} | TLR complex |Toll-like receptor signalling pathway ^{ns}
-	Black (R3)	-	10.8 (172, -25.7)	Anatomical structure development | Actin binding | Extracellular matrix | NA
-	Purple (R11)	-	7.32 (68, -13.4)	Skeletal system development |Extracellular matrix structural component | Extracellular region part | NA
-	Red (R4)	-	5.44 (173, -9.7)	Ossification^{ns} | Growth factor activity^{ns}| Organelle^{ns} | Hedgehog signaling pathway
Red (H6)	-	-	-	Interphase of mitotic cell cycle^{ns} | oxidoreductase activity ^{ns} | Cell fraction^{ns} | Cell Cycle^{ns}



The **C3** module is the rat (**R6, ***) and human (**H5, •**) consensus module ($\tilde{\chi}$-score=15.5, $\log_{10}$ *p*-value=-52.8) with a strong association with the alginate culture trait (cor=0.44, *p*=8e-10) and an functional annotation relating to cytokine activity.

Figure 5.15: Circos plot to present rat ('**R**' prefix) and human ('**H**' prefix) module overlaps with consensus modules ('**C**' prefix). Figure derived from tabular data where rows (rat or human modules) and columns (consensus modules) are represented by coloured segments (inner circle) the size of which defines the total number of genes that overlap with the rat or human modules. Ribbons connect rows and columns and are coloured by consensus module to show the overlap with each rat or human module. The outer two rings define the relative contribution of each cell in a table to the row and column totals (stacked bar plots). The figure summarises the tabular data shown in **Figures 5.13, 5.14** and **Table 5.4**.

5.3.5: Differential eigengene network analysis shows greater preservation of network structure across species than between rat Affymetrix and Illumina data

Differential eigengene network analysis was used to define the strength of the correlation preservation for all eigengene pairs across the two networks, **Figure 5.16**. There was strong evidence for eigengene network preservation across rat and human networks (D = 0.9), which was greater than in preliminary analysis greater undertaken between the rat Affymetrix and Illumina data sets (**4.3.5**).

Two consensus modules (C4 and C5) were highly correlated between the two data sources and formed a 'meta-module'. Meta-modules can represent biologically relevant gene 'super-sets'. The meta-module contained 134 genes, which was functionally annotated by the terms: 'immune system process' (biological process, p=2.6e-15), 'plasma membrane' (cellular compartment, p=2.3e-13) and 'signal transducer activity' (metabolic function, p=4.7e-11).

Species-specific modules were defined by considering the overlap with consensus network modules. Only one module from the human meta-set was not represented in the consensus network (H6), but this had no functionally enriched annotation, **Figure 5.14** and **Table 5.4**, and demonstrated the most overlap with the grey module of unassigned genes. As such this was not considered to be biologically meaningful.

There were a number of rat meta-matrix specific modules for which there was no eigengene representation in the consensus network, **Figure 5.13** and **Table 5.4**. As a consensus module can only be created if a module exists in both data sets this may represent modules that are specific to the rat meta-matrix rather than species-specific *per se*. **Figure 5.15** highlights orphan modules.

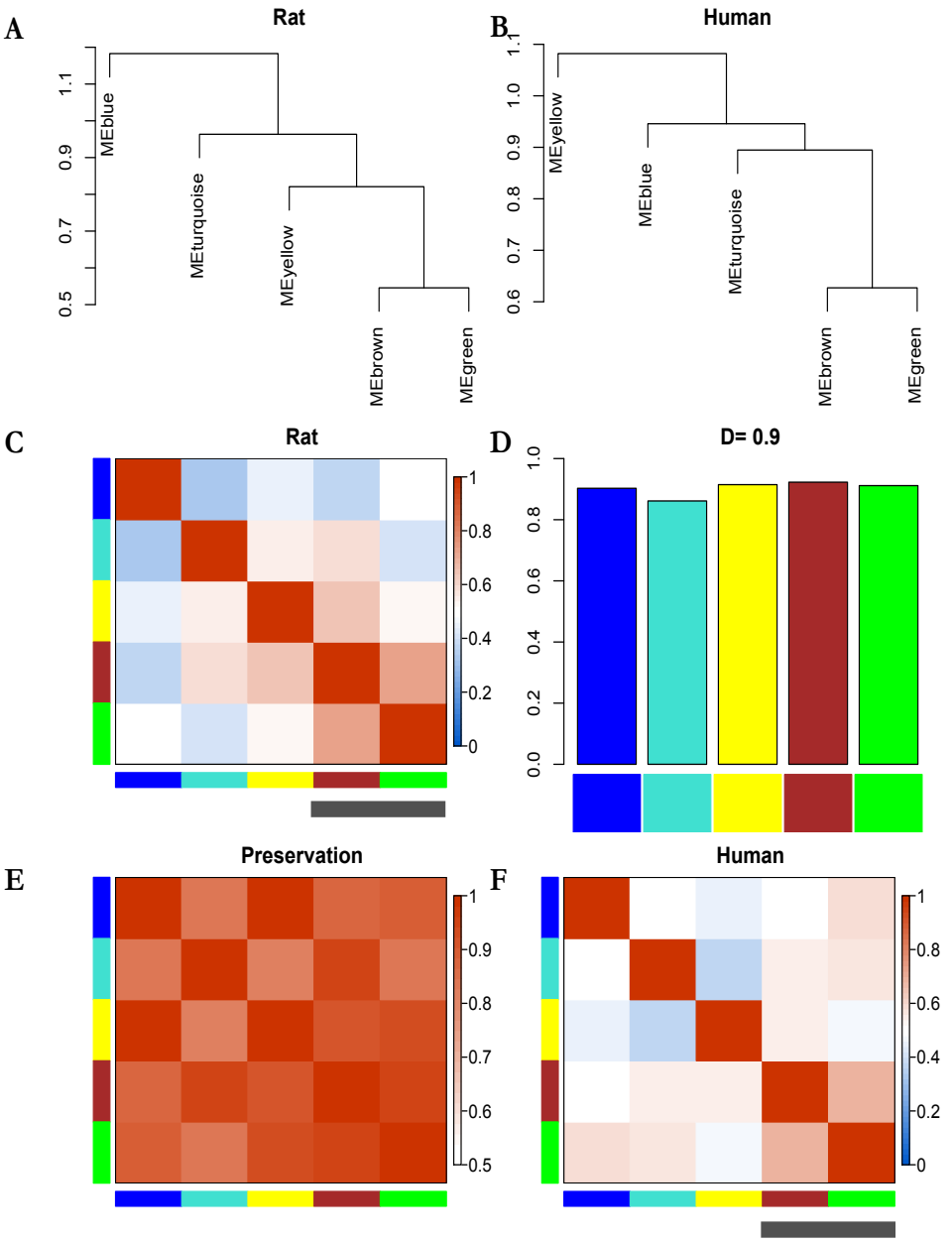


Figure 5.16: Differential eigengene network analysis across rat and human networks using Affymetrix array data. **A.-B**. Clustering dendrograms of consensus module eigengenes demonstrates the presence of a meta-module consisting of the green and brown consensus modules (black bar). **C**.-**F**. Heat-map plots of eigengene adjacencies for each of the eigengene networks (**C**- Rat; **F** – Human). Each of the rows and columns indicate an eigengene labelled by the consensus module colour. Red indicates high-adjacency (positive correlation), whilst blue indicates the inverse, as depicted by the colour legend. **D**. Barplot of the preservation of the consensus eigengene relationships between the two meta-sets. Additionally, the overall network preservation measure *D* for this differential analysis is also provided. **E**. Adjacency heatmaps for the pair-wise preservation networks of the two meta-sets. Each consensus module is represented by the rows and columns with the level of red saturation indicating adjacency according to the colour legend.

5.3.6: Relating phenotypic traits to modules

The categorical membership of samples to phenotypic traits was assessed for correlation with eigengene modules for each species-specific network. In order to reduce the number of phenotypic terms, and to explore emerging trends in the data, samples descriptors were iteratively merged into larger trait groups consisting of broader, more inclusive terms. Binary classification tables were produced to categorise the arrays based upon disease status, tissue source, location, or experimental condition.

Rat module-trait relationships

The strongest module-trait relationship was found for the R4 module, **Figure 5.17** – the combined traits of 'hypertrophic' and 'transitional' for chondrocytes derived from growth-plate studies were positively correlated with this module eigengene (cor = 0.71, p = 4e-27), which was enriched for the KEGG canonical 'Hedgehog-signalling' pathway; this module was negatively correlated with the physiological trait for 'epiphyseal', 'proliferative' and 'resting' growth plate zones. The R6 module eigengene was positively correlated with the term 'perturbed', an aggregated group consisting of surgical and inflammatory models of osteoarthritis in the rat, in addition to three-dimensional culture models (alginate and fibrin) and healing ligament samples. In contrast, the 'monolayer' trait was negatively correlated with R6 (cor = -0.55, p = 6e-15)

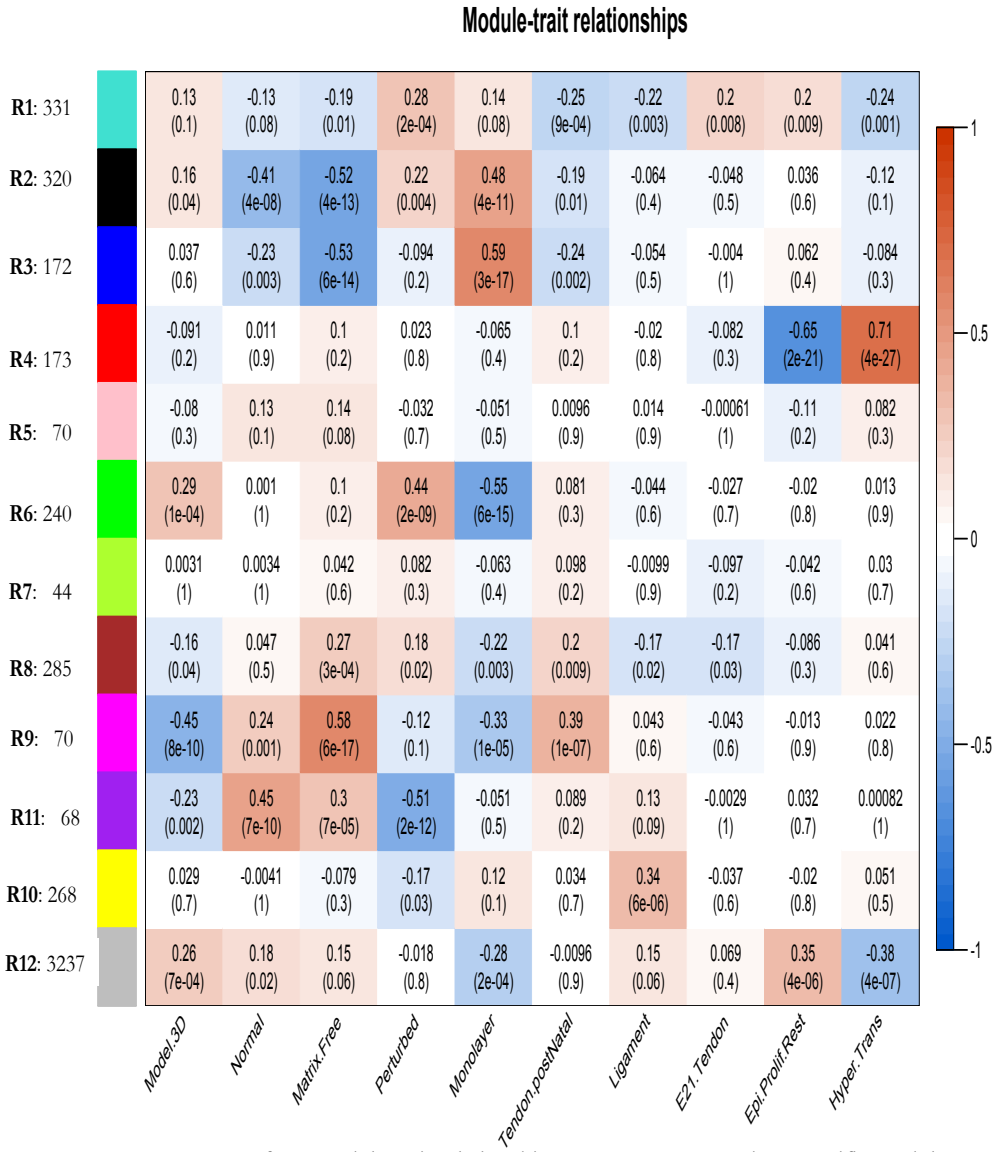


Figure 5.17: Heat-map of **rat** module-trait relationships - Rows represented rat-specific module eigengenes, columns are descriptive terms based upon the data sample descriptions. Samples may appear in more than one descriptor. Cells are coloured on the strength of the correlation between the module and the trait (side bar). The absolute correlation is provided in the cell with the associated *p*-value in parentheses below. Strong positive correlations are seen between the R4 module and chondrocytes in transitional/hypertrophic zones of growth plates, also between the R3 module and cells in monolayer. Strong negative correlations are seen between the R4 module and the remaining growth plates zones, suggesting a biologically relevant difference in the expression of these module genes. Normal cartilage ('normal') is negatively correlated with the R3 and R2 modules, found to be associated with monolayer culture.

Human module-trait relationships

Initially twelve phenotypic categories for the human meta-set samples were chosen based upon the declared sample phenotypes in the microarray repositories. On this basis, however, there was no phenotypic trait that demonstrated a strong positive association with any of the module eigengenes. To aid definition of sample traits in the human data a multi-dimensional scaling plot was used to define divergent sample clusters (data not shown). This allowed samples to be grouped for module-trait correlation in a manner that was not intuitive from an understanding of the published sample phenotypes alone, for example, by disease status.

The strongest module-trait correlation was seen with the human module H5 and 'Group 2' (cor = 0.57, p = 1e-15), **Figure 5.18**, which consisted of: foetal cartilage, chondrocytes in alginate beads treated with conditioned media from fibroblasts from patients with rheumatoid arthritis (RA), whole joint tissue at 17 weeks and differentiating mesenchymal stem cells at week 1 and 2. This module was negatively correlated with samples assigned to 'Group 1', which consisted of chondrocyte condensations, and other alginate cultures treated with pharmaceutical small-molecules that inhibit inflammatory process, e.g. corticosteroids and non-steroidal drugs.

Across the two species networks the R6 and H5 modules correlated with chondrocytes within inflammatory environments, but also differentiating chondrocytes and mesenchymal stem cells and chondrocytes in three-dimensional alginate constructs. When these samples ('perturbed' (rat) and 'Group 2' (human)) were associated with the consensus network C3 module there was also a moderate correlation, C3 (cor = 0.44, p=8e-10), **Figure 5.19**, indicating that the consensus

module, representative of the R6 and H5 modules, was also positively correlated with an inflammatory or differentiating phenotype in chondrocytes.

Hub genes are conserved between rat and human module

As a measure of the validity of defining hub genes from the consensus module for each gene associated with the R6 or H5 modules, the Gene Significance (GS) for the trait 'perturbed' was plotted against intra-modular connectivity. For both species-specific networks there was a moderate, but significant, correlation between these two measures, range: cor = 0.36 – 0.38, p=2.2e-05 - 9.4e-09, **Figure 5.20**, comparable in both the rat and human data. Using the Module Membership measure kME in the rat was plotted against kME in the human. This presented strong evidence that module-hub genes were conserved in the R6 and H5 modules, cor = 0.49, p = 6.7e-16, **Figure 5.21**.

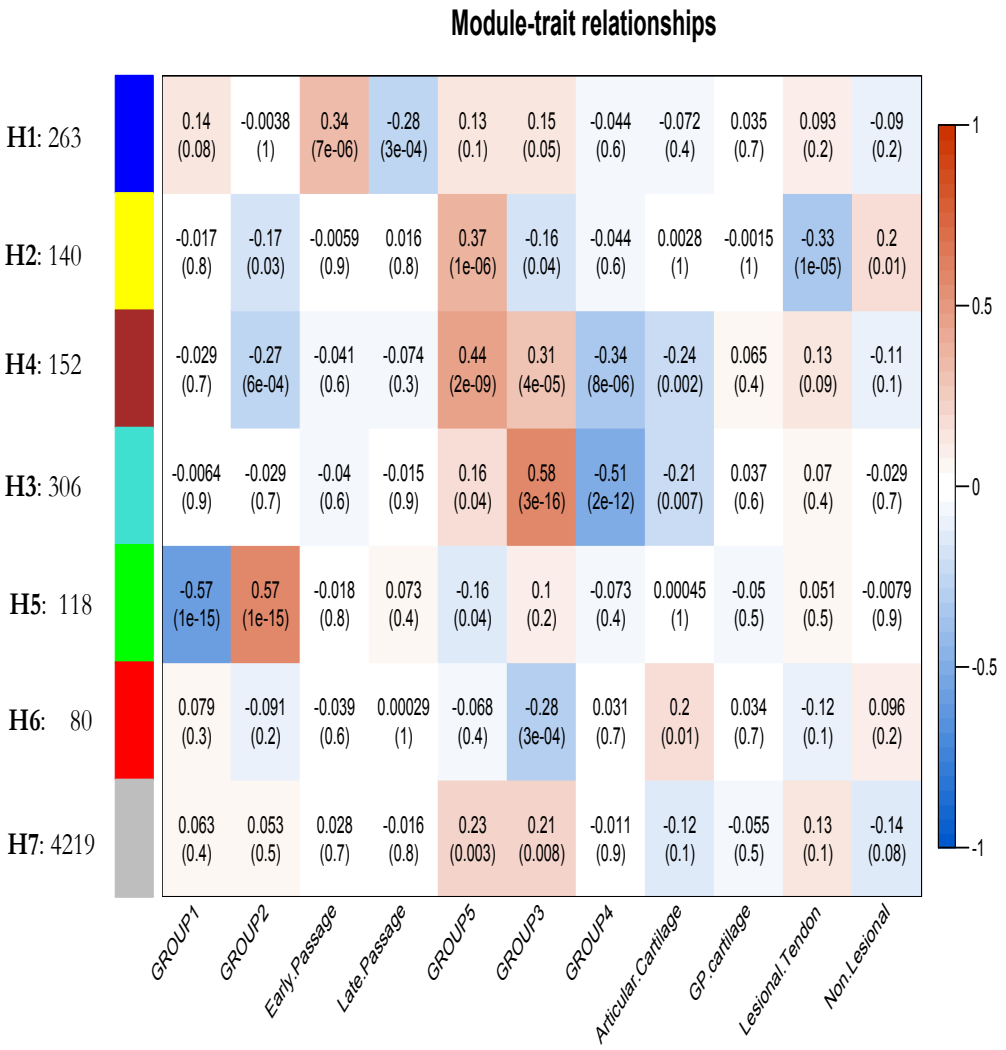
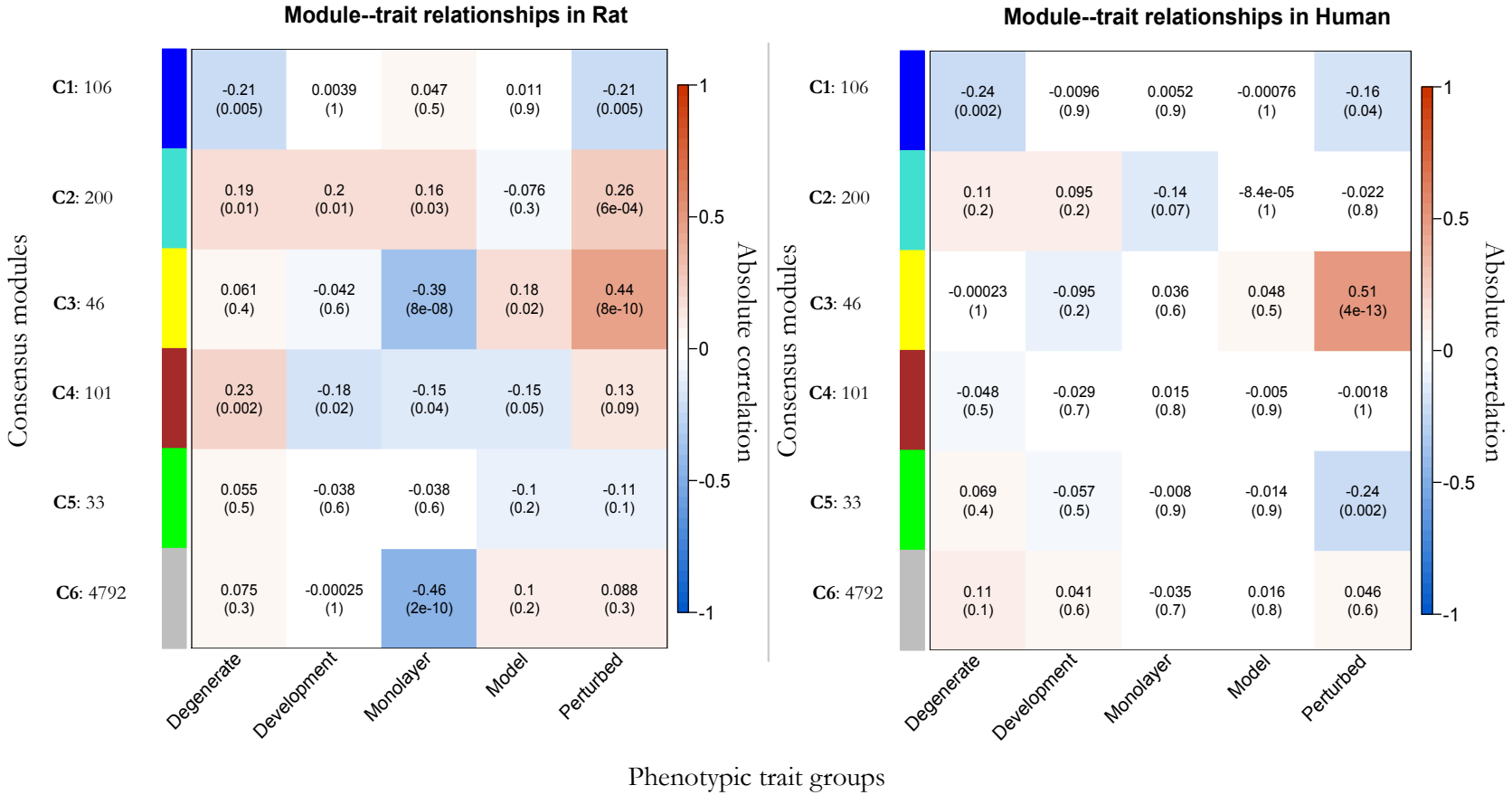


Figure 5.18: Heat-map of **human** module-trait relationships - Rows represented meta-set module eigengenes, columns are descriptive terms based upon the data sample descriptions. Samples may appear in more than one descriptor. Cells are coloured on the strength of the correlation between the module and the trait (linear heatmap). The absolute correlation is provided in the cell with the associated *p*-value in parentheses below.

Figure 5.19: Module:trait associations - Using the consensus eigengene modules the yellow module (**C3**), rows, was the only module that shared a positive correlation with a trait (columns) across the rat and human meta-sets (defined by coloured vertical bar). Use a conservative method to define a union statistic the lowest correlation, where directionality was shared, leaves the C3 module with a consensus correlation= 0.44 (p=8e-10) for the trait 'perturbed'. This trait consisted of all the samples from the human 'Group 2' and the rat 'perturbed' definitions.



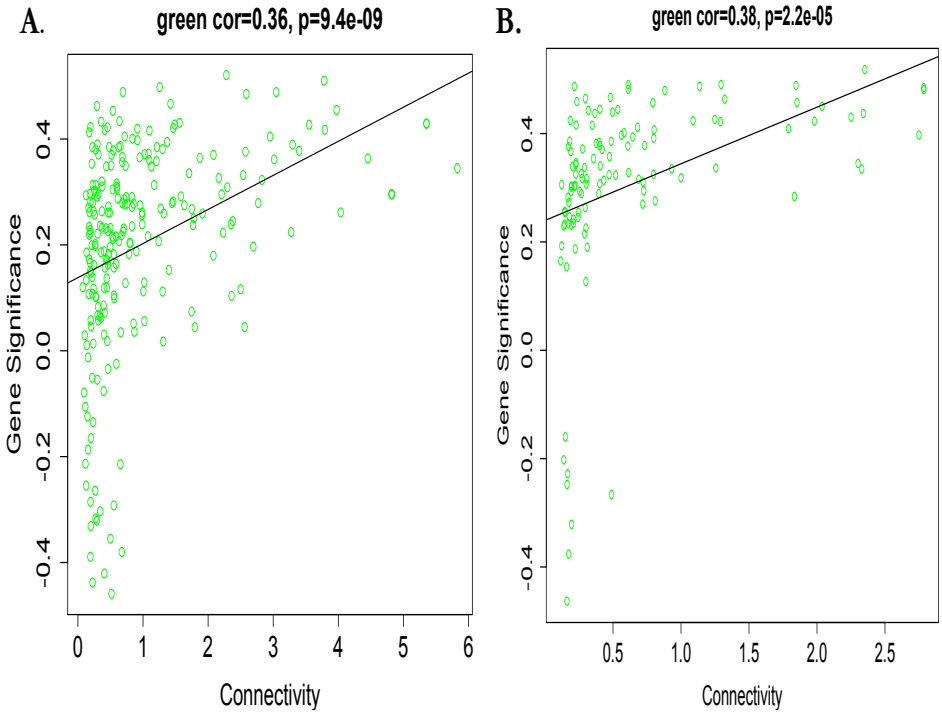


Figure 5.20: Gene significance for the consensus trait 'perturbed' shows moderate correlation with intra-modular connectivity in the rat (**A**) and human (**B**). Although the correlation is not strong there results are comparable across the two networks indicating that hub genes are also likely to have strong phenotypic associations

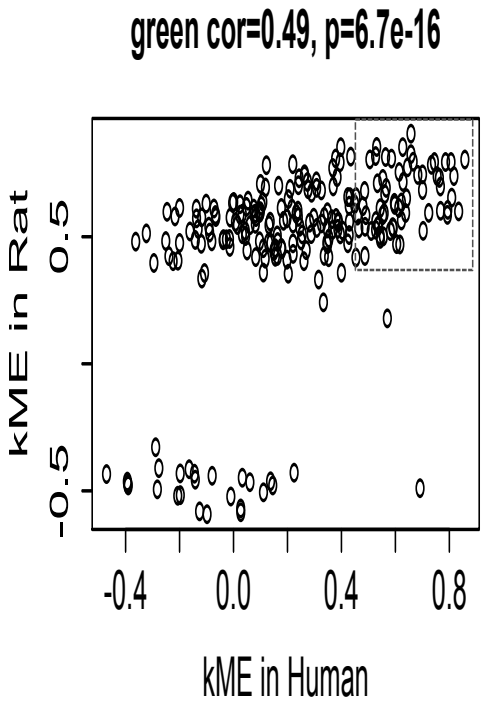


Figure 5.21: Only genes with a positive association for module membership (*k*ME values) were plotted, per module, for the rat (y-axis) and human (x-axis) networks using the rat modules as the reference data set. This provided a visual assessment of hub gene conservation. The rat R6 module *k*ME is correlated with the human *k*ME indicating that genes with high module membership in the rat module are are likely to have high module membership in the human module (box).

5.3.7: Differences in human and rat network modules define osteoarthritis-associated module

There were several rat modules for which there was no equivalent module in the consensus network or any evidence of overlap with consensus modules, suggesting that these modules were not present in the human network. The R11 module was defined by the gene ontology terms 'skeletal system development' and 'extracellular matrix structural component' and was moderately associated with normal cartilage samples, **Figure 5.17**. The R11 module had low module preservation statistics when compared to the human network, **Table 5.4**. Using genes within the R11 module with a *k*ME >0.6 (46 genes) this module was significantly associated with the terms 'skeletal system development' (p = 4.4e-6), 'extracellular region part' (p = 3.9e-4) and 'osteoarthritis' (p = 9.4e-3) using the Genetic Association database annotations in DAVID. A number of genes with robust associations with osteoarthritis were present in this module: *Col2a1*, *Frzb*, *Tnfrsf11b*, along with key regulators of chondrogenesis and cartilage turnover, *Wif1*, *Dlk1* and *Scrg1*. The genes with the highest correlation with the skeletal development module were: *Mfge8*, *Chst3*, *Eps8l2*, *Mif2* and *Col9a2*.

5.3.8: Defining conserved hub genes between rat and human modules

H5 and R6 consensus hub genes are key pro-inflammatory mediators

In order to assess the conservation of modules the gene Module Membership, as described previously, was the qualifier of module hubs. Those genes in paired-modules that had the highest *k*ME represented hubs in both rat and human networks. The top 20 genes for each module, for which there was a consensus match, are presented in **Table 5.5.** These module hubs had highly significant gene ontology annotations given the small input number and also had known protein-

protein interactions, **Figure 5.22**. When the network topography of the R6 and H5 modules was considered the most highly inter-connected genes differed between the two modules, **Figure 5.23**, which may reflect biologically relevant differences between the two data sets.

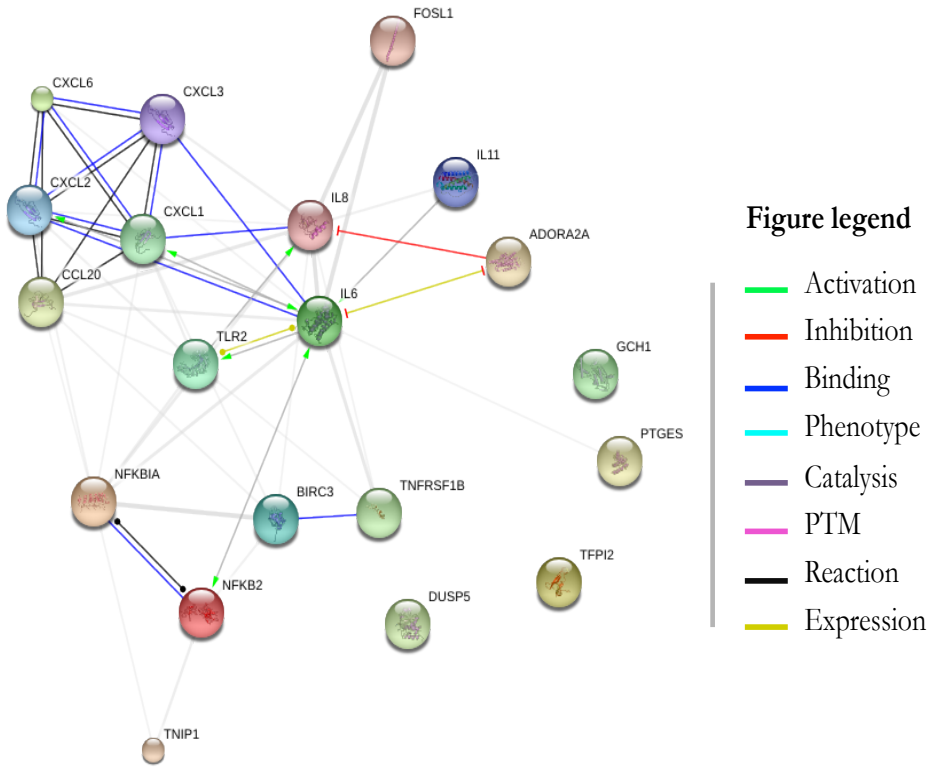


Figure 5.22: Consensus module hubs for rat and human R6 and H5 modules presented with protein-protein interactions using STRING. The top 20 genes with the highest ranking *k*ME were chosen from each of R6 and H5. Legend colours define protein-protein interactions.

Figure 5.23:
Cytoscape generated connectivity networks for the R6 and H5 modules using the genes with the highest module membership across both species.

Intra-modular connectivity (**figure legend**) of the consensus hub genes differs between the rat and human networks. In the rat module (top) the chemokines *Ptges* and *Cxcl5*, and *Upp1* were the most highly connected (figure legend, Node degree).

In the human module the network topography was such that *IL6*, *IL11*, and *MMP10* were the most highly connected. *CCL20* and *CXCL1* were moderately connected in both networks.

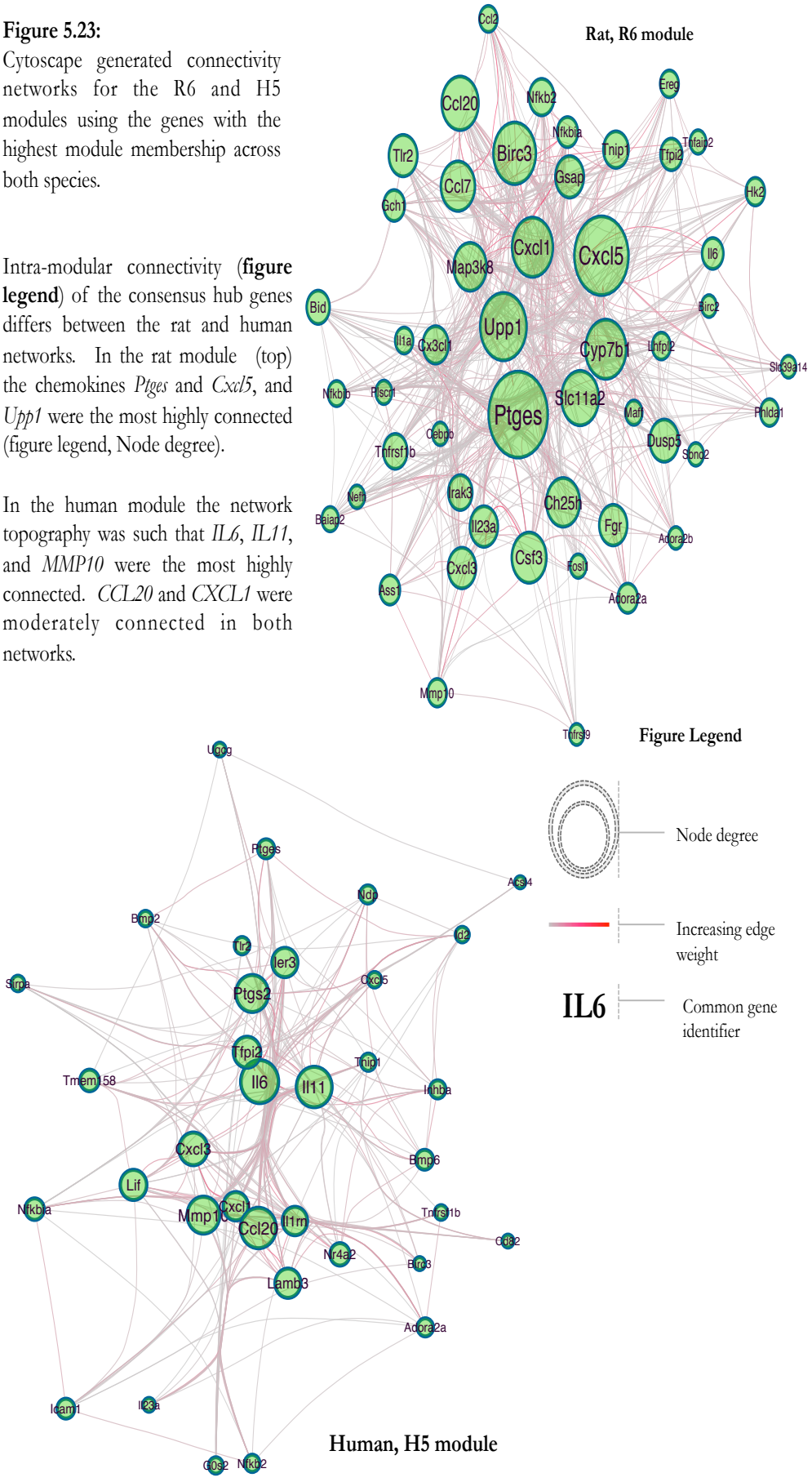


Table 5.5: Top 20 consensus hubs between human and rat modules – consensus modules are shown for reference, **Table 5.4**. Gene ontology functional annotations based upon only the top 20 hub genes. The H5 and R6 modules are both associated with the perturbed chondrocyte trait. Gene ontology annotations: BP – biological process; CC: cellular compartment; MF: metabolic function; KEGG: canonical signalling pathway.

Human Module	Rat Module	Consensus Module	Top 20 consensus hubs	Gene Ontology BP | MF | CC | KEGG pathway (Bonferroni p-value <0.05)
Yellow (H2)	Yellow (R10)	Blue (C1)	ACTN2, APOBEC2, COX6A2, CSRP3, HSPB3, ITGB1BP2, MYF6, MYH7, MYOZ2, NRAP, PITX2, PKIA, PPP1R3A, SLC25A4, SMPX, TNNC1, TNNI1, TNNT1, TRDN, KLHL41	Muscle contraction | Cytoskeletal protein binding | Sarcomere | NA
Blue (H1)	Turquoise (R1)	Turquoise (C2)	MCM2, KIF18B, FOXM1, PBK, BUB1, C12ORF48, ASPM, TTK, CCNB1, KIF2C, CDK1, GINS1, CDCA8, KIF11, MCM5, MELK, KIF23, SPAG5, AURKB, CCNA2	Nuclear division | ATP binding | Microtubule cytoskeleton | Cell cycle
Green (H5)	**Green (R6)**	**Yellow (C3)**	NFKBIA, CCL20, IL11, PTGES, TFPI2, FOSL1, CXCL1, BIRC3, TNIP1, CXCL6, DUSP5, TLR2, CXCL3, ADORA2A, TNFRSF1B, NFKB2, IL6, CXCL2, GCH1, IL8	Defense response | Chemokine activity | Extracellular region part | Cytokine-cytokine receptor interaction
Turquoise (H3)	Brown (R8)	Brown (C4)	GPR116, EMCN, GPR4, TIE1, PLVAP, PODXL, CD93, APLNR, MYCT1, RASIP1, CALCRL, TRPC6, PDE2A, PRKCH, F11R, C1ORF115, MMRN2, MFNG, NPY1R, LPAR6	Immune system process | Receptor activity^{ns} | Integral to plasma membrane | Leukocyte transendothelial migration
Brown (H4)	Magenta (R9)	Green (C5)	IGSF6, IL10RA, PTPRC, LCP1, NCKAP1L, HCK, CD53, TBXAS1, VAV1, LAPTM5, CCR1, PLEK, CYBB, ITGB2, CSF1R, ITGAM, FCER1G, NCF4, LCP2, HLA-DMB	Signal transduction^{ns} | Receptor activity | Plasma membrane | Neuroactive ligand-receptor interaction

5.3.9: Hierarchical clustering discriminates samples by consensus hub genes

To explore species-associated similarities in the module hubs the gene expression profile of the twenty consensus hub genes defined for the H5 and R6 module were considered for all rat and human samples. Samples from human data contained within the trait collection 'Group 2', in general, demonstrated higher expression of the twenty consensus hub genes, **Figure 5.24**. In particular chondrocytes in alginate beads, treated with synovial fibroblast conditioned media derived from rheumatoid arthritis patients, were the most extreme.

Similarly, in the rat, hierarchical clustering defined alginate bead cultures and 'perturbed' sample as showing higher expression of the top twenty consensus hub genes, **Figure 5.25**. Using unsupervised hierarchical clustering alone it was suggestive that this gene profile could discriminate alginate three-dimensional cultures and the 'perturbed' chondrocyte phenotype from other samples. Monolayer samples, in the rat study, had the lowest expression of the consensus hubs, whilst native chondrocytes and tenocytes had an expression profile that lay between these extremes. Notably, expression profiles from surgical and inflammatory models of osteoarthritis in the rat were not as extreme as those from three-dimensional cultures.

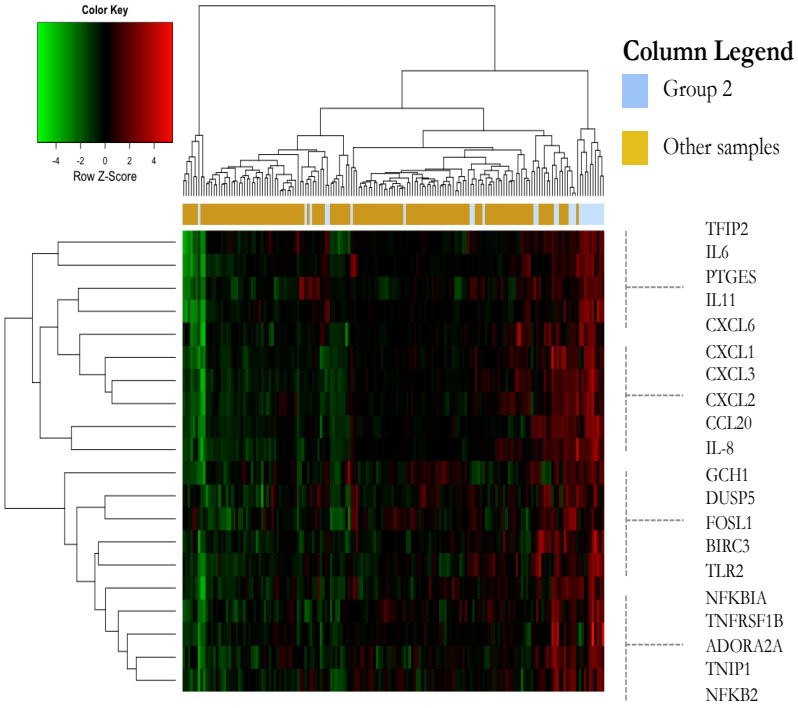


Figure 5.24: Heatmap – Genes expression values for the top 20 consensus hubs defined for the **H5** and **R6** modules are plotted for all 166 **human** meta-set samples. The samples defined within this group are coded as described in the **column legend**. High expression is defined by red cells and low expression by green cells. Rows represent genes (row annotation). Hierarchical clustering does not completely discriminate between groups, however, the majority of Group 2 samples were defined within clades showing higher expression of these hub genes.

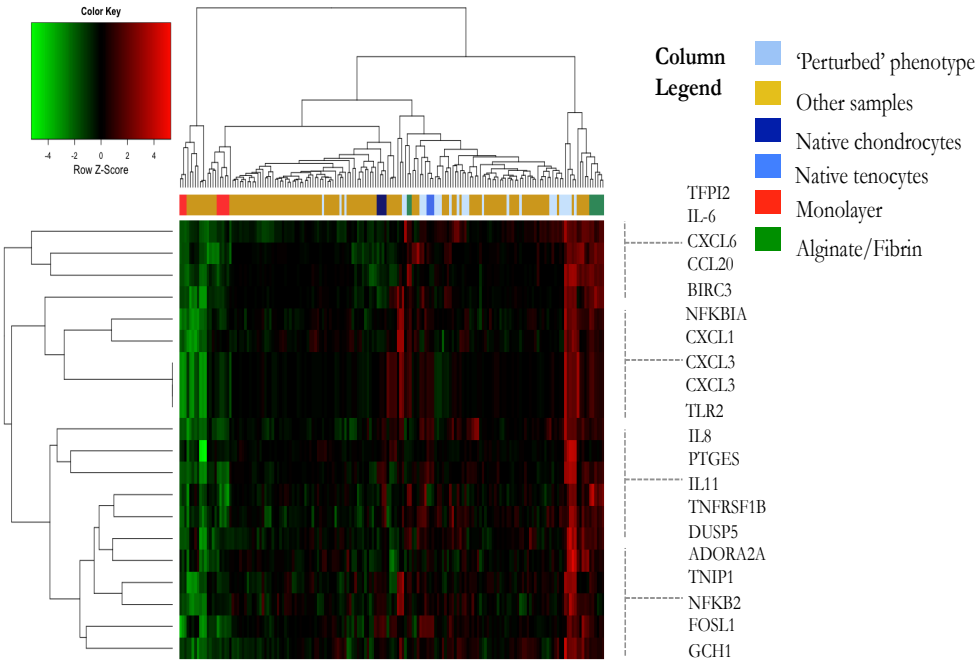


Figure 5.25: Heatmap – Genes expression values for the top 20 consensus hubs defined for the **H5** and **R6** modules are plotted for all 170 **rat** meta-set samples. Legends as for **Figure 5.24**. Alginate and fibrin three-dimensional cultures represent the most extreme expression profile in the dataset and show the highest expression of the consensus hub genes.

5.3.10: *IL-6* containing inflammatory profile is predictive of three-dimensional cultures

A key objective in this analysis was to establish a gene cohort that would accurately and consistently predict class membership for three-dimensional culture conditions based on a limited gene expression profile. Feature selection for class prediction was based on the C3 module (n = 46 genes) that was shown to have a strong association with the gene expression profile of chondrocytes in alginate beads.

Class prediction, by a least-shrunken centroids approach, was considered for membership of either model cultures (alginate and fibrin) or any other sample (tissue, monolayer). For rat samples, using a 13 gene signature, low classification error rates were achieved using the training data set (0.026 – 0.046 on five iterations), **Figure 5.26**. The top predictors were *Cxcl6*, *Lif*, *Ccl20*, *Il11* and *Il6*. On test data the model gene signature correctly classified all other samples as 'not model' (posterior probability (PP) = 0.61-0.99), however, for two alginate cultures in the test set the posterior probability was less confident (PP=0.42-0.59).

The same process applied to the human data was able to discriminate alginate bead cultures, with or without conditioned media, from other samples (total error rate: 0.013-0.027 on five iterations), **Figure 5.27**. Between 13 and 16 genes achieved low error rates,

however, 13 genes performed better. The prediction model differed to the rat with *TFPI2*, *IL6* and *PTGS2* the best predictors; *LIF* and *NR4A3* were not present in the human signature, whereas *MMP10* was included. When considered against the network topology, **Figure 5.23**, the human class predictors were

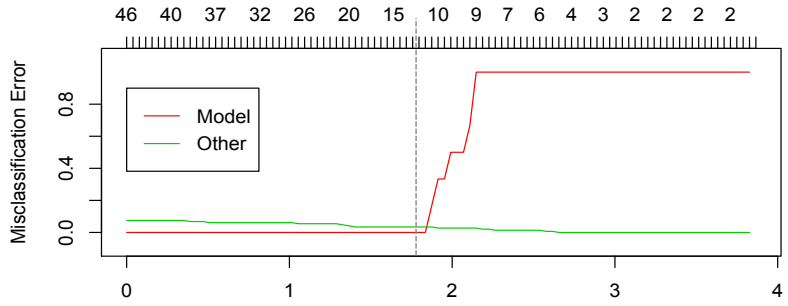
consistent with the most highly connected nodes in the C3 module; this did not appear as consistent for the rat data. On test data the human signature performed well with all models correctly identified (PP = 0.56-0.79) and all other samples classified as non-model cultures (PP = 0.75-0.99).

To assess how well each gene signature would perform across species each signature was used to define the test data set from the other species. The human signature *versus* the rat test data was poorly predictive with both alginate cultures in the test data set misclassified (PP = 0.713). In contrast, the rat signature *versus* the human test set correctly classified all three alginate samples as model cultures (PP = 0.6-0.89). Using the rat gene signature did reduce from 77.3% to 70% the proportion of the data variation described in principal component analysis (data not shown).

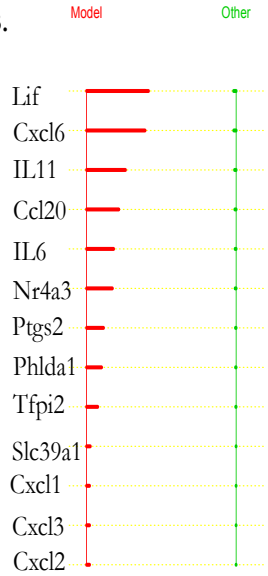
A fundamental test for any prediction model is how well it deals with unseen data, especially where small class sizes can result in all members of class being within the training set and so biasing the data. The Illlumina data set (**Chapter 2**) was not used in training and served as an independent data set. Of the thirteen genes identified as the rat gene signature interleukin 11 (IL-11) probes could not be identified from the Illumina RatRef.v1 probe manifest file or from re-annotation of probes by Ensembl (November 2014), therefore, only twelve gene were used to discriminate between the Illumina samples. A low classification error rate was evident for the binary classification of 'models' versus all other sample (0.033), indicating that this gene cohort was a sufficient model of three-dimensional culture models in two independent data sets, **Figure 5.28**.

Figure 5.26: Class prediction for alginate cultures from **rat** gene expression data

A.



B.



A: Plot of error rate (y-axis) against threshold value, *Delta* (lower x-axis). As the number of remaining genes (upper x-axis) from C3 module decreases the misclassification error for model cultures is low. At Delta=1.796 (vertical line), 13 genes (**B**), the smallest number of errors are made(4)

B: Plot of shrunken centroids for each class for genes surviving the threshold for at least one class

C: 2x2 confusion matrix to define the number of samples in each class that were correctly identified. No model cultures were misclassified in the training set as 'other', whilst five samples known not to be model cultures were classified as such. Overall error rate = 0.033

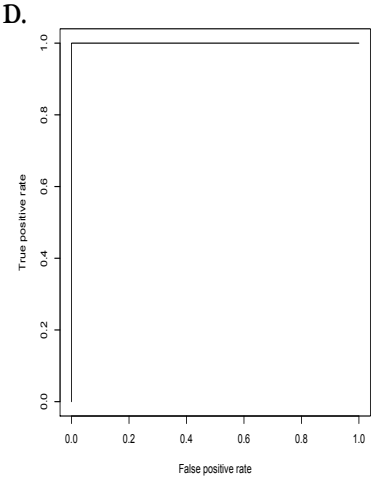
C. **Prediction on genes from consensus module**

Training	Model	Other	Error Rate
Model	6	0	0.0
Other	5	142	0.034

D: Receiver-operator curve: False positive rate (x-axis), True positive rate (y-axis) – area under the curve (AUC) = 1.

E: Principal component analysis (n=170) rat expression data on 13 genes from **B**. First two components explain 73.9% of the variation in the data.

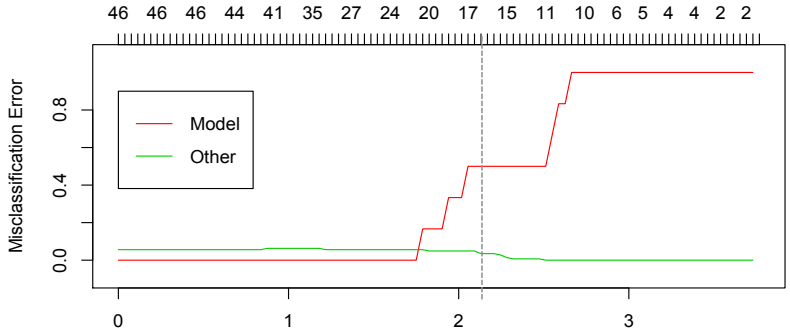
D.



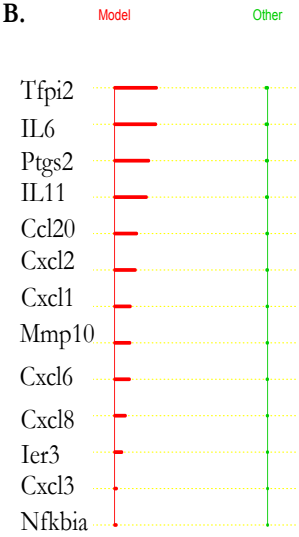
E.

Figure 5.27: Class prediction for alginate cultures from **human** gene expression data

A.



B.



A: Plot of error rate (y-axis) against threshold value, *Delta* (lower x-axis). As the number of remaining genes (upper x-axis) from C3 module decreases the misclassification error for model cultures is low. At Delta=2.394 (vertical line), 13 genes (**B**), the smallest number of errors are made(2)

B: Plot of shrunken centroids for each class for genes surviving the threshold for at least one class.

C: 2x2 confusion matrix to define the number of samples in each class that were correctly identified. Three model cultures were misclassified as 'other' in the training set, whilst one sample known not to be model cultures were classified as such. Overall error rate = 0.027

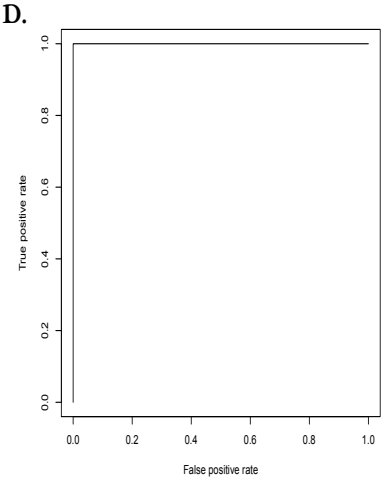
C. **Prediction on genes from consensus module**

Training	Model	Other	Error Rate
Model	3	3	0.5
Other	1	142	0.006

D: Receiver-operator curve: False positive rate (x-axis), True positive rate (y-axis) – area under the curve (AUC) = 1.

E: Principal component analysis (n=166) human expression data on 13 genes from **B**. First two components explain 77.3% of the variation in the data.

D.



E.

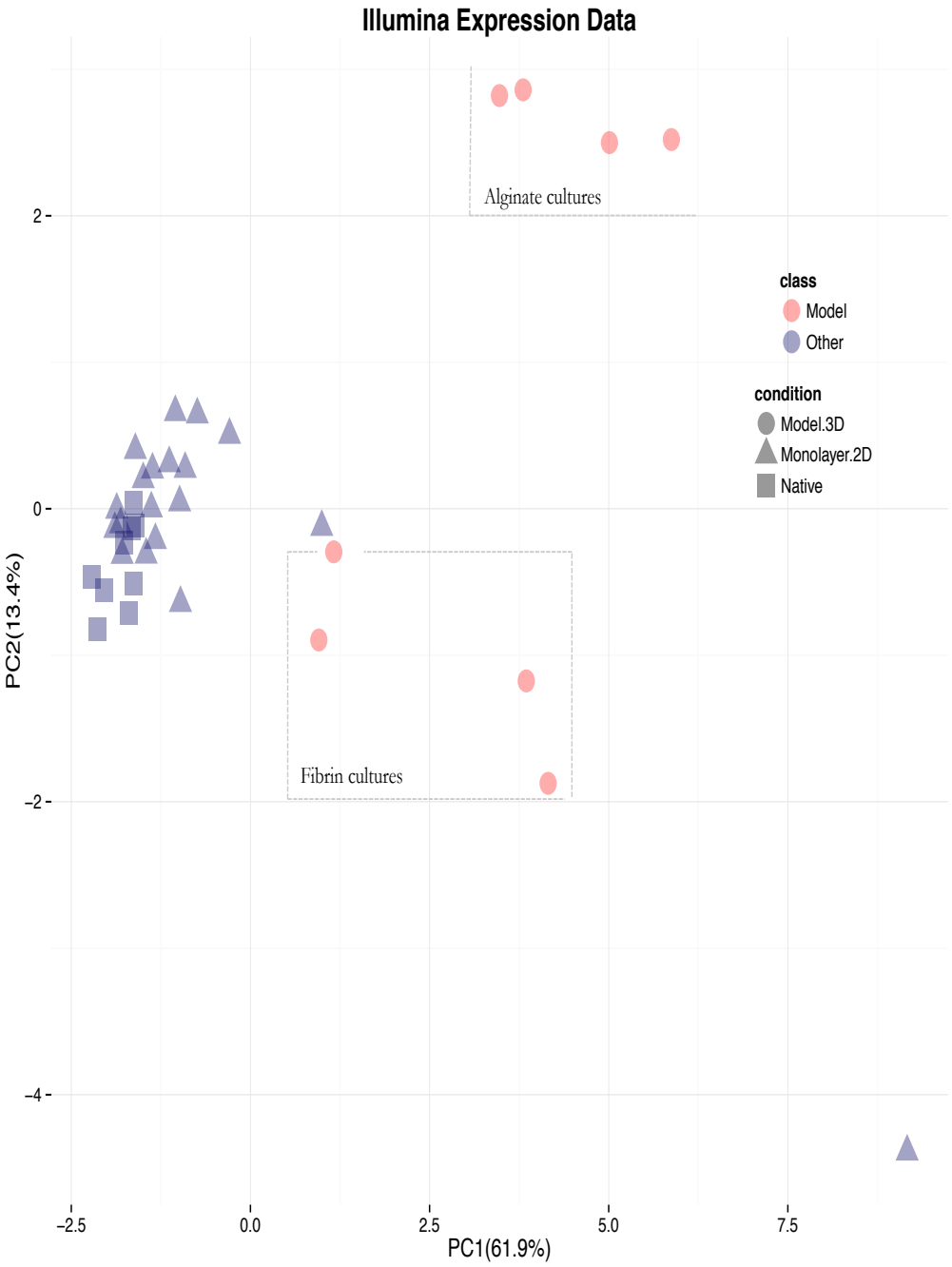


Figure 5.28: Principal component analysis – Loess normalised expression data from Illumina data (Chapter 2, n=36) comprising twelve genes (excluding *IL11*) from the rat class predictor model. Data points represent individual arrays and are coded as per the figure legend. Although alginate and fibrin cultures do not cluster together they are distinct from both native and monolayer samples on the basis of expression of twelve genes. Two samples, marked as triangles, do not follow this trend completely and represent samples from dermal fibroblasts at passage five.

Prediction on 12 genes from consensus module

Training	Model (7)	Other (28)	Error Rate
Model	7	0	0.0
Other	1	22	0.043

Overall error rate: 0.033

In the majority of cases three-dimensional model cultures are correctly identified using a 12 gene signature.

5.4.11: Alginate culture-associated module is not preserved in human liver transcriptome network

Rationale

To challenge the methodology, and condition or cell specificity of module definitions a gene expression meta-matrix from human liver microarray studies (n=150) was prepared and the co-expression network compared to those of rat and human cartilage/tendon data. These microarrays were included in the meta-matrix without *a priori* understanding of liver studies. Specifically only data derived from the Affymetrix Human Genome U133 Plus 2 array was chosen. The same gene-set applied to the rat and human comparison (n=5278) was used so that the modules would be comparable. Liver data was selected because there was a number of large data sets available using a single platform.

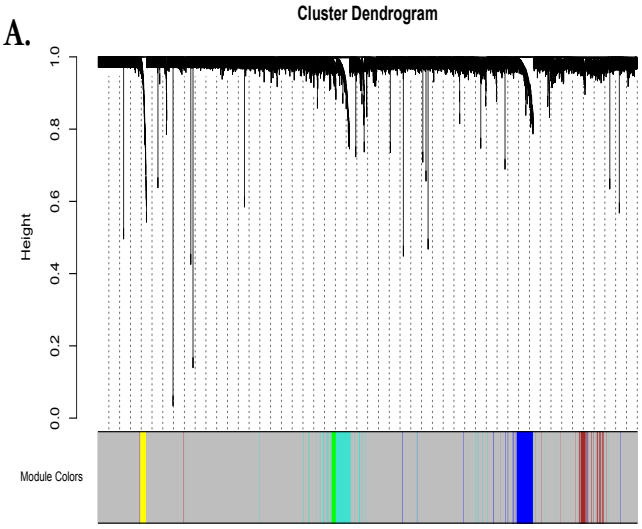
Rat and human *IL-6*-containing module is poorly preserved in a liver co-expression network

A consensus network was prepared from the human cartilage/tendon and liver networks, **Figure 5.29**. High overall preservation of consensus eigengene networks (D = 0.93) between human cartilage/tendon and liver networks was demonstrable, **Figure 5.30**. High preservation of tissue networks within a species is not unexpected. Specifically the eigengene network contained a meta-module, an aggregation of highly correlated eigengenes, which was strongly preserved in both the cartilage/tendon and liver networks. Considering whether the human cartilage/tendon network structure was well described in the consensus network the gene overlap in modules was investigated. Only four genes from the H5 module were found any consensus module, other than the unassigned module, **Figure 5.29**.

A consensus network was produced using the rat cartilage/tendon and human liver networks to considered whether the R6 module would also be excluded from a consensus network structure with human liver data. Analysis of the eigengene network structure, found the lowest network eigengene preservation, D=0.86, of these analyses (data not shown). Comparing the rat set-specific network structure against the consensus network the R5 had 46 genes that overlapped with a consensus module, **Figure 5.31**, i.e the total number of genes in the R5 module.

Finally, for both the human and rat cartilage/tendon networks a permutation test was prepared to quantify how preserved set-specific modules were in the consensus network prepared with the human liver data. In both cases the R6 and H5, *IL-6*-containing modules, were the lowest scoring modules for preservation in each consensus network, **Figure 5.29** and **5.31**. This is supportive of the earlier findings of the specific association of the R6 and H5 modules with a perturbed or differentiating chondrocyte phenotype.

A.



B.

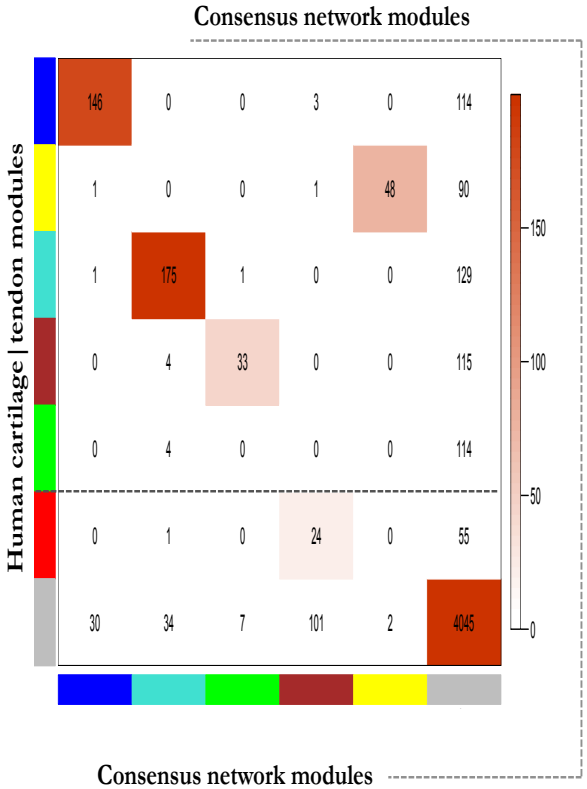


Figure 5.29:
Gene dendrogram for human liver and cartilage|tendon meta-sets with consensus module definition. Five modules are defined by the consensus network (**A**).

The overlap of genes within modules derived from the cartilage|tendon meta-set (rows) were compared to those from the consensus network (columns) (**B**). All human modules overlapped with the consensus modules except for the human H5 module (dashed lines). Cells contain the number of genes common to both modules. Colour intensity (vertical colour chart) represents the number of overlapping genes

C.

Cartilage| Tendon Module	Module size	Z-score (summary)	Log10-p (Bonferroni summary)
blue	263	30.34	-264.17
turquoise	306	25.15	-165.94
yellow	140	13.19	-39.39
grey	400	9.97	-27.65
red	80	9.34	-20.82
brown	152	6.46	-9.65
green	118	3.17	-2.41
gold	100	2.85	-1.83

C. Module preservation table for human cartilage|tendon modules compared to the consensus network module definitions. The human H5 module (green) has a Z-score comparable to that of the 'gold' module, which consists of a random allocation of genes. Z-scores lower than 5 are unlikely to be preserved.

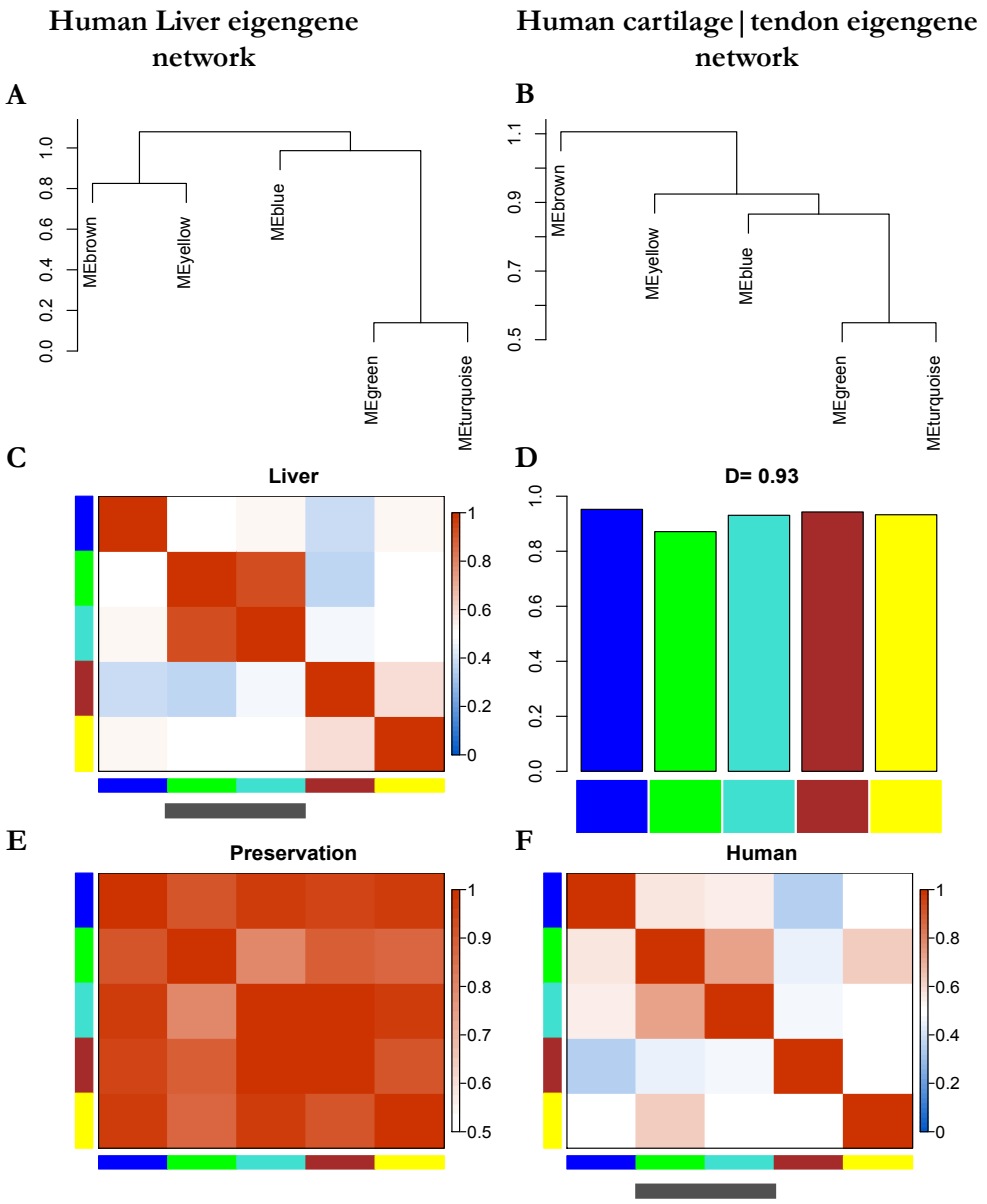


Figure 5.30: Differential eigengene network analysis across **human** liver and cartilage/tendon networks using Affymetrix array data. **A.-B**. Clustering dendrograms of consensus module eigengenes demonstrates the presence of a meta-module consisting of the green and turquoise consensus modules. **C**.-**F**. Heat-map plots of eigengene adjacencies for each of the eigengene networks (**C**- Liver; **F** – Human cartilage/tendon). Meta-module is defined by black bar in each plot.. Each of the rows and columns indicate an eigengene labeled by the consensus module colour. Red indicates high-adjacency (positive correlation), whilst blue indicates the inverse, as depicted by the colour legend. **D**. Barplot of the preservation of the consensus eigengene relationships between the two meta-sets. Additionally, the overall network preservation measure *D* for this differential analysis is also provided. **E**. Adjacency heatmaps for the pair-wise preservation networks of the two meta-sets. Each consensus module is represented by the rows and columns with the level of red saturation indicating adjacency according to the colour legend. High preservation is often evident between tissues from the same species.

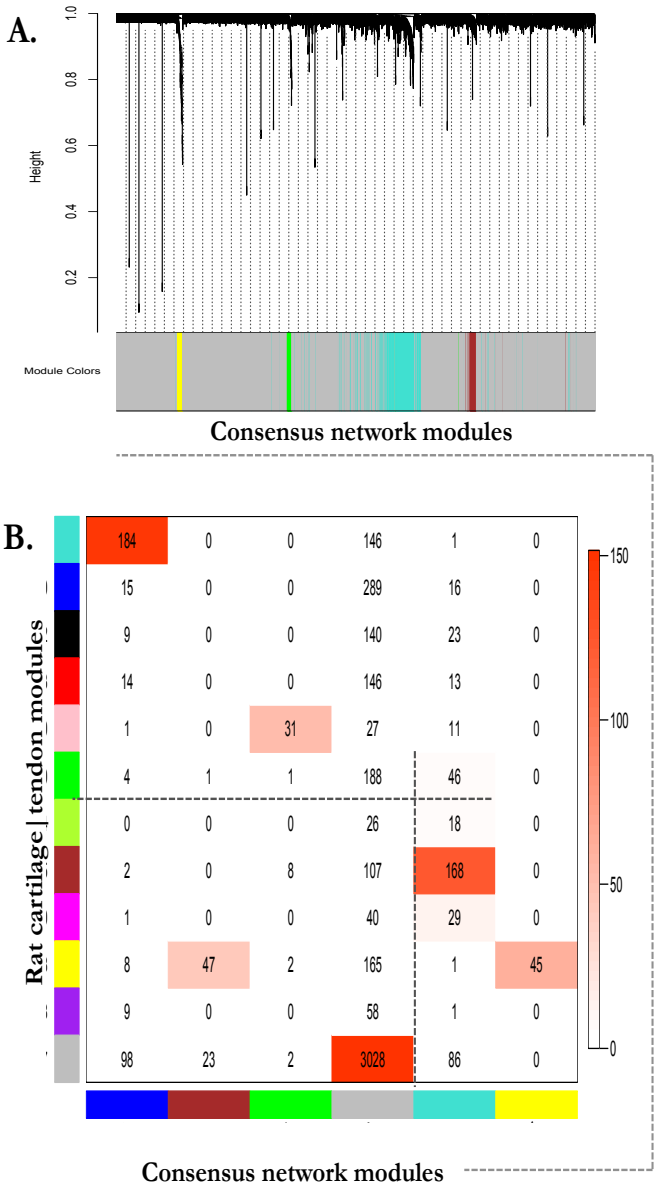


Figure 5.31 : Gene dendrogram for human liver and rat cartilage|tendon networks with consensus module definition. Five modules are defined by the consensus network (**A**).

The overlap of genes within modules derived from the rat cartilage/tendon network (rows) were compared to those from the consensus network (columns) (**B**). Only four consensus modules (plus the unassigned grey module) overlapped strongly with those modules specified in the rat network (**Figure 5.13**). Specifically the green, **R5**, module had 46 genes that overlapped with the turquoise module. Matrix cells contain the number of genes common to both modules. Colour intensity (vertical colour chart) represents the degree of overlap.

C.

Rat Modules	Module size	Z-score (summary)	Log-p (Bonferroni summary)
brown	285	23.85	-159.85
turquoise	331	23.93	-154.23
greenyellow	44	11.61	-53.29
yellow	268	8.66	-21.85
pink	70	8.83	-18.20
black	172	8.41	-15.64
blue	320	6.68	-13.42
green	**240**	**4.76**	**-5.62**
magenta	70	4.81	-5.32
gold	100	3.83	-3.53
grey	400	2.12	-1.01
red	173	1.49	-0.38
purple	68	-0.86	0

C: Module preservation table for rat cartilage/tendon modules compared to the consensus network module definitions. The rat **R5** module (green) has a Z-score comparable to that of the 'gold' module, which consists of a random allocation of genes. Z-scores lower than 5 are unlikely to be preserved, i.e. the R5 module is not preserved in the human liver network.

5.4: Discussion

As demonstrated in **Chapter 4** weighted co-expression network analysis may be used to compare modular units of transcriptome networks to extract highly co-expressed aggregates of genes that may confer deeper understanding of the regulatory mechanisms underlying a particular phenotype. Global network comparisons may be made between conditions, or, as is the case in this study, between species. This approach, where multiple small gene expression data sets are merged and compared between species, has been well-described for the mouse and human brain transcriptome (Miller, Horvath et al. 2010).

The remit of this study was to employ this methodology to define divergent and conserved gene co-expression modules in rat and human transcriptomic networks derived from cartilage or tendon, and their derivative cells. At the outset the goals of this study w

ere four-fold: i) to integrate microarray gene expression profiles in cartilage and tendon and compare this across species; ii) highlight data gaps to facilitate development within the musculoskeletal research community for future targeted research projects; iii) define *in-vitro* culture models relative to normal and disease-associated gene expression profiles; iv) to consider whether rat studies represent adequate proxies for complex musculoskeletal disease in the human. Each point shall be addressed in turn in four sections.

5.4.1 Implementation of data integration methods for gene expression profiles

In trying to compare the results of differential expression analysis in **Chapters 2** and **3** it became evident that placing the findings in a wider context simply by

matching genes or functional annotations across publications would not be sufficient. For example, findings of functional annotations associated with muscle in tendon expression profiles could not be validated by analysis of published gene lists. Meta-analysis and data integration methods were explored to identify the most effective methods to compare gene expression studies and utilize complete sample transcriptome profiles.

Sequence- and target-matching probes across unique platforms

As described in the introduction, reports in the literature indicate that sequence-matching of probes from multiple platforms can result in more robust and repeatable estimates of gene expression. The exercise in this study was driven by the poor availability of microarray studies profiling cartilage and tendon in public repositories both for rat and humans and the desire to maximize the number of studies that could be included. The low number of probes that could be confidently sequence-matched and also represented unique targets would have prohibited the application of the co-expression network analysis. Although the final analysis was restricted to platforms from the same manufacturer (Affymetrix) several platform generations were used in both the rat and human studies. Evidence suggests that reproducibility of analysis and improved correlation between samples may be achieved through sequence- and/or target-matching of probes between Affymetrix generations (Nimgaonkar, Sanoudou et al. 2003, Hwang, Kong et al. 2004, Elo, Lahti et al. 2005). It may be recommended that sequence- and target-matching of probe-sets across Affymetrix generations is undertaken in further rounds of analysis or where data is added.

It is clear from this analysis that noise in the individual studies cannot be resolved by global normalisation techniques. Furthermore, addition of data did not facilitate the creation of discrete phenotypic groups, rather adding data contributed to noise in the analysis effectively 'filling in' the expression space. Of the studies included information on temporal batching of expression profiles would have facilitated batch-correction of each data set prior to integration; this is not currently a requirement of the MIAME recommendations. The most popular cross-platform normalisation strategies require equivalent covariates to be present in samples across multiple platforms, consequently, they cannot be relied upon to be able to consistently separate data arising from diverse sources or experiemental techniques in a biologically relevant way. Hence, the simplest z-score normalisation strategy was applied as it was the least disruptive to the underlying biological information.

Whether z-score normalisation was the most appropriate is worth discussion. In a study normalizing across five Affymetrix platform generations and 6,926 experiments Autio, *et al* (2009) reported that their Array Generation based gene Centering (ABC) normalisation technique outperformed standardization, equalized quantile, and Weibull distribution based methods (Autio, Kilpinen et al. 2009). This methodology is not implemented in *R* and there is little evidence for its widespread use elsewhere; furthermore, z-score standardization retained clustering of biological replicates, whilst overcoming within-platform clustering. The decision to employ z-score normalisation was supported by data from Andreas, *et al* (2009), from which the human alginate bead data arose, where the distinct

separation of samples treated with different pharmaceutical compounds shown in the publication was retained (Andreas, Häupl et al. 2009). This implies that simple z-score normalisation across Affymetrix generations did not adversely impact the data distribution or sample relationships.

Limitations in available musculoskeletal gene expression profiles governs study inclusion criteria

Data integration, whether using quantitative clinical data or, in this study, microarray expression data, is bounded by the availability of good quality data sets. The dearth of microarray studies, relative to other tissue sources, for both cartilage and tendon is a primary limitation of this study. Comparable cross-species analysis (Miller, Horvath et al. 2010) using the WGCNA methodology used much larger datasets (>1000 microarrays from approximately 20 individual studies) with very much higher gene expression correlations across the tissue samples.

As a consequence of this it may be argued that the inclusion criteria of this study was too broad, including as it did both cartilage and tendon, monolayer, novel cell culture techniques and interventional studies. Important associations between co-expression modules and sample phenotypes may have been obscured by the diverse conditions and would have benefitted from a more restricted inclusion criteria, for example, only whole cartilage. In response, the results outlined in **Chapter 2** and **3** highlighted the importance of tissue comparisons in musculoskeletal studies especially where the aim of the investigation is the derivation of novel tissue biomarkers or an understanding of common regulatory mechanisms. The number of available datasets is critical to the use of methods using co-expression data (Ostlund and Sonnhammer 2014). As demonstrated in

Chapter 3 small datasets do not have a stable network structure and so attempts were made to meet the recommended numbers for this type of analysis.

The diversity of studies, use of alternative breeds and genders, lack of tissue controls within many studies and the use of multiple microarray platforms make data integration especially problematic, if not incomparable. Strategies were used to best deal with these qualitative issues. By using only raw expression data it was possible to apply pre-processing and normalisation strategies consistently and universally. Stringent quality control of each study data set ensured that outlier arrays were removed, despite these having been included in the original publications. Using a co-expression analysis method avoided integration of differential expression statistics across studies, which are often not reproducible (Ostlund and Sonnhammer 2014), by considering only the global network structures.

In summary, the strategies applied here were driven by the data availability and the methodology could be refined where more specific data sets become available. Despite these limitations a systematic data integration approach was developed with results derived from careful quality control of studies and consistent manipulation of raw expression data, approaches that should lend credence to the results presented.

5.4.2 Global comparison of the rat and human transcriptome networks should inform future musculoskeletal research

This analysis has highlighted some of the key limitations to developing good models for musculoskeletal disorders and obstacles to the extrapolation of results from model species to humans. Firstly, there is a lack of publically available data,

especially for tendon from either species. Secondly, the quality of these data sets was often poor, with retention of outlier arrays, small group sizes and re-use of single biological replicates across multiple arrays. Thirdly, when placed in a wider context most gene expression profiles for cartilage and tendon clustered together indicative of the noisy nature of the data. For some studies it was not possible to replicate the published clustering of experimental groups even where comparable bioinformatic techniques were used. Together these points suggest that concerted community efforts are required to standardize approaches to gene expression profiling of cartilage and tendon. Whilst collecting sufficient disparate data sets to perform co-expression analysis may be a promiscuous approach this analysis represents a first step in developing a systems understanding of cartilage and tendon responses to perturbations and is a novel resource for musculoskeletal researchers to exploit for future experimental design.

Validity of cross-species comparisons of expression data

Low, but positive, correlation between the expression profiles of the rat and human cartilage and tendon meta-sets was shown, which still permitted comparison of these data sets as defined by documentation associated with this method. The cross-species expression correlations defined by Miller, *et al* (Miller, Horvath et al. 2010) between mouse and human brain transcriptomes were highly positive, however, in that study all data sets were derived from whole tissue transcriptome analysis. In contrast this study datasets were derived from diverse sources and multiple Affymetrix platforms. Confidence in this analysis of these data sets is derived from the very high module preservation across the species networks and strong consensus eigengene network structure.

Studies considering the comparative structure of the rodent and human transcriptomes have found that tissue-specific transcriptome architecture is highly conserved across mouse, rat and human with tissue-specific variance accountable for the greatest alterations in expression profile, more so than either perturbations or disease (Prasad, Kumar et al. 2013). It would perhaps be expected that the expression correlation would have been higher between the two meta-matrices. The compound nature of the meta-matrices, containing two tissues, and multiple array platforms may have influenced this correlation. Furthermore, the inclusion of heterogenous cell populations from tissue, as compared to the monoculture *in vitro*, can influence the generation of co-expression modules (Gaiteri, Ding et al. 2014). Random resampling of these data sets, however, did not greatly alter the intra-species expression correlation, nor the network structures, and so the consensus network architecture generated is likely to be robust to the inclusion of other data sets. It is critical that these findings are validated against new gene expression profiles.

'Normal' cartilage module is not represented in the human network

Through differential analysis of the network architecture the presence of a rat-specific module, associated with unperturbed cartilage samples, was found. Within this module a number of genes with strong osteoarthritis associations, and novel candidates, were presented. This module, annotated with terms relating to cartilage and skeletal development, was strongly correlated with the normal cartilage phenotype in the rat network, specifically with samples derived from controls; it was strongly negatively correlated with the 'perturbed' phenotype, i.e. those from surgical interventions. Given the role of these genes in cartilage health it is not initially intuitive as to why this module is so poorly preserved in the

human network. The rat meta-matrix contains a number of normal cartilage control samples and sham interventions that comprise the 'normal' phenotype; there is no equivalent for this in the human meta-matrix. This highlights the dearth of control data available in human studies against which to compare the rodent models.

This module contained genes recently found to be highly differentially expressed between paired normal and OA-affected samples (Ramos, den Hollander et al. 2014) including *Frzb*, *Tnfrsf11b* (which encodes the protein osteoprotegerin) and *Col9a1*. Furthermore, key modulators of cartilage turnover, *Wif1* (Witte, Dokas et al. 2009, Stock, Böhm et al. 2013) and regulators of differentiation, *Dlk1* (Chen, Qanie et al. 2011) and *Scrg1* (Ochi, Derfoul et al. 2006). The most highly correlated genes were those with relative sparse associations with cartilage – *Mfge8*/lactadherin has previously been shown to be expressed in cartilage (Yoshimi, Miyaishi et al. 2005), but only recently reconsidered. In ColIX$^{-/-}$ 21 days post-natal mouse cartilage increased MFGE8 protein expression was demonstrated (Brachvogel, Zaucke et al. 2013). It is not clear from this analysis whether these findings are a consequence of a rat-specific functional module or reflect bias in the data as a consequence of few normal cartilage samples in the human meta-set.

Relevance of animal models of musculoskeletal disease

Ultimately it is expected that rodent model studies are translatable and inform human disease pathogenesis in some way. The gene expression studies presented in **Chapters 2** and **3** may, as with many other rodent model expression studies, have restricted translational potential to human musculoskeletal disorders and there is little evidence in the literature to suggest that the validity of assumptions

have been investigated. A literature search of rat studies found that these studies are difficult to compare in isolation as they are derived from different disease models, tissue sources, *in vitro* interventions, breeds and sexes, and are often underpowered.

In general the transcriptome networks for both species were comparable, however a number of rat modules were not present in the human network. Furthermore, the connectivity of genes within a preserved module was not equivalent across the species and the components of the alginate bead culture gene signature also differed between rat and human networks. Human gene expression studies of whole cartilage or tendon were scant; only three microarray samples in the human data set were defined as coming from ostensibly normal cartilage (Minogue, Richardson et al. 2010). There were no normal tendon gene expression profiles available in the public repositories.

Given the dearth of data, and differences in network structure, no assumptions can be made about the relevance of rodent gene expression profiles from musculoskeletal disease models for informing the complex human condition.

5.4.3: Conflicting roles of IL-6

In this study a co-expression module consisting of multiple chemokines was demonstrated, most notably represented by members of the IL-6 family of chemokines: *Il-6*, *Il-11* and *Lif*. A pleiotrophic battery of functions are associated with classical and *trans*-signalling by IL-6 including the transition from innate to acquired immunity in inflammation, and roles in regeneration, metabolic control and bone metabolism/osteoclast differentiation (Scheller, Chalaris et al. 2011). This indicates both physiological and pathological roles for IL-6 signalling. This

is evident in the samples from the rat and human data sets showing the greatest correlation with this module. In the rat, expression profiles arising from osteoarthritis models and three-dimensional culture models were associated with the same consensus module as human data arising from foetal cartilage, differentiating MSCs and alginate cultures.

There is conflicting evidence for the association between IL-6 and osteoarthritis. IL-6 and its receptor are found in the synovial fluid of symptomatic and osteoarthritic joints (Doss, Menard et al. 2007, Tsuchida, Beekhuizen et al. 2012), yet both anabolic and catabolic effects of Il-6 have been found. In the work of de Hooge, *et al* (2005) male *Il-6* $^{-/-}$ mice developed evidence of osteoarthritis more rapidly that age-matched wild-type controls; this was characterized by complete cartilage erosion, subchondral bone sclerosis and ossification of the collateral ligaments of the femoro-tibial joint (de Hooge, van de Loo et al. 2005). This conflict is further highlighted by evidence both catabolic effects of Il-6, through induction of MMP3 and MMP13 production, and anabolic roles through the induction of alpha1-anti-trypsin and TIMP1. Ryu, *et al* (2011) demonstrated that IL-6 was a potent effector of *Epas1*/HIF2α-induced cartilage destruction (Ryu, Yang et al. 2011). Hypoxia-inducible factor 2-alpha (HIF2α) has been previously demonstrated to be a regulator of cartilage catabolism (Yang, Kim et al. 2010); inhibition of IL-6 resulted in reduced MMP3 and MMP13 production in response to intra-articular HIF2α adenovirus. Referring to the discussion in **Chapter 3**, Litherland, *et al* (2008) reported cartilage collagenolysis dependent on PI-3Ksignalling by the activation of Akt by IL-6 stimulation (Litherland, Dixon et al. 2008). These findings are relevant with respect to the evidence presented in this thesis of high *Il-6*, *Mmp13* and *Hif1α* expression in rat chondrocytes in alginate

beads (**3.3.3**) and the associated prediction of active PI-3K signalling by pathway topology analysis (**3.3.7**).

These findings in rat chondrocytes would suggest a catabolic process in action in alginate cultures. However, in a study comparing IL-6 production by chondrocytes from symptomatic cartilage defects, osteoarthritic or normal cartilage Tsuchida, *et al* (2012) found that chondrocytes retained in a regeneration culture system of collagen type II-coated filters produced higher levels of IL-6, by ELISA, than the synovial fluid from each associated donor category (Tsuchida, Beekhuizen et al. 2012). Levels of IL-6 were highest in chondrocytes from OA cartilage. Although antibody-mediated inhibition of IL-6 in regeneration cultures had no effect on cartilage matrix production the addition of recombinant IL-6 and IL-6R resulted in significant increases in GAG production in chondrocytes derived from healthy donors and reduced production, but not content, of GAGs in OA-derived chondrocytes. Other co-culture studies have indicated a negative modulation of matrix by IL-6 (Leyh, Seitz et al. 2014a, Leyh, Seitz et al. 2014b).

Studies are also conflicted with regard to genetic susceptibility associated with *IL-6* polymorphisms. Susceptibility to hip and knee OA associated with *IL-6* promoter polymorphisms could not be demonstrated (Valdes, Arden et al. 2010), however, a more recent study observed a high frequency of *IL-6* polymorphisms associated with knee and hand OA, but not hip OA (Cai, Sun et al. 2014).

5.4.4: Alginate culture associated with inflammatory profiles

This thesis has not sought to develop novel organotypic culture systems rather interrogate the gene expression profile of common *in vitro* models and question the rational for their use with the view that this will inform future developments. The

analysis here confirmed the findings of enhanced expression of genes associated with inflammatory profiles in these models. This was preserved in studies using human chondrocytes in a comparable three-dimensional culture model. Furthermore, when restricted to twenty genes these models had an expression profile more extreme than diseased cartilage or tendon in the human study. To interpret these findings alternative solutions should be considered.

Alginate, or alginic acid, is a seaweed-derived linear polysaccharide consisting of alpha-L-guluronic acid (G) and beta-D-mannuronic acid (M). It is considered to contain impurities including lipopolysaccharide (LPS, endotoxin). Alginate alone has also been shown to induce the production of pro-inflammatory cytokines, including IL-6, in macrophages through NF-κB activation (Yang and Jones 2009).

Basal expression and secretion of IL-6 has been found in OA chondrocytes retained in alginate beads, as a function of culture duration, with an increase in expression following IL-1β application and variable reductions in expression in response to the application of common NSAIDs (Sanchez, Mateus et al. 2002).

Given the critical role of pro-inflammatory mediators in our understanding of cartilage and tendon pathology stringency in cell culture preparation cannot be overlooked. Screening for endogenous endotoxin is not routine outside of GMP labs and those familiar with biomaterial research. Also, as the role of IL-6 is conflicted this needs to be separated from any inherent pro-inflammatory and endotoxin sequestration effects of biomaterials (Breger, Lyle et al. 2009). Anecdotal reports of reduced viability of chondrocytes in alginate beads and the evidence of this in **Chapter 2** indicate another argument for investigating pro-inflammatory effects more closely (**2.3.8**). Further work should include batch-

testing of alginate for endotoxin prior to use, in addition to a time-course study of IL-6 production in both alginate and fibrin three-dimensional culture models.

The findings from the rat Affymetrix and Illumina class prediction were most comparable with a study by Andreas, *et al* (2009) (Andreas, Häupl et al. 2009). It was found that a subset of samples consisting of chondrocytes suspended in alginate beads and exposed to RA serum had a profile very like that of the rat alginate culture studies. In the study by Andreas, *et al*, healthy adult chondrocytes, initially expanded in monolayer, were encapsulated in alginate beads in methodology comparable to the study presented here. Alginate beads were exposed to conditioned media from cultured synovial fibroblasts from RA patients or normal controls, and further small molecule interventions were also performed. There is remarkable overlap between the differential gene expression profile of chondrocytes in RA fibroblast conditioned media in this publication and the class prediction gene cohort presented in this study. Additionally, the clustering of samples, including therapeutic interventions, is comparable to the original publication supporting the assertion made in this study that biologically relevant data distribution is preserved after z-score normalisation.

5.4.5: Biological relevance of study

Whilst functional and histo-morphological equivalence with native tissue may represent the ultimate goal for biological tissue engineering many studies often rely on a limited panel of established genes to define success. This approach can neglect the complex gene interaction mechanisms diminishing the significance of the analysis (Haynes, Higdon et al. 2013). Whilst large transcriptome surveys and co-expression analysis are not practical in most cases the profiling of a validated gene signature for class prediction could be a systematic screening procedure to

demonstrate progression of a novel organo-typic culture method. In other terms, the expression space within which these culture systems reside needs to be delimited. Using a linear analogy differentiated tissue and dedifferentiated monolayer cells represented either extreme of a gene expression spectrum in this study; defining where along that line a novel culture system 'sits' by way of screening of modules of highly co-expressed genes may allow higher throughput testing of novel systems and provide rational regulatory targets for perturbation or therapeutic interventions.

In this chapter a cohort of genes is defined that may represent a class prediction signature capable of discriminating between standard three-dimensional culture systems and monolayer and native tissue sample in the rat and human. This signature contains a number of IL-6 family members and other chemokines with known musculoskeletal disease and development associations. Reluctance to make more confident statements about the significance of this finding stems from the small number of samples relative to the whole study and, therefore, statistical concerns regarding bias. This signature would require validation, either by qPCR or microarray, using the same three conditions across multiple culture time-points. Concomitant protein validation would also be required.

The tissue-specific architecture of transcriptomes is highly conserved across mouse, rat and human studies (Prasad, Kumar et al. 2013), however, this was not unambiguously demonstrated for cartilage and tendon alone in this study. The authors of this study noted that comparative studies of mouse, rat, and human transcriptomes in response to perturbations, were still missing. This type of investigation would require the use of comparable tissue sources and identical experimental conditions. The work presented in this chapter, to the author's

knowledge, is the first to consider transcriptomic responses to perturbations arising both in cartilage and tendon.

This study makes reference to the small number of relevant cartilage and tendon studies made available to researchers both for rats and humans. Only through the permissive inclusion policy of this study was it possible to gather a sufficient number of data sets required for network co-expression analysis. Clearly the use of matched tissue and experimental conditions for rat and human would be ideal, but this type of data is not currently available in public repositories.

Little is understood of the comparability of rat gene regulatory networks relative to humans in complex disease models. It is imperative, therefore, that there is community consensus on the nature of animal models for osteoarthritis or tendinopathy rather than small 'first-in-field' studies that are difficult to translate into complex human musculoskeletal disease.

References

Allen, J., S. Wang, M. Chen, L. Girard, J. Minna, Y. Xie and G. Xiao (2012). "Probe mapping across multiple microarray platforms." Briefings in Bioinformatics **13**(5): 547-554.

Andreas, K., T. Häupl, C. Lübke, J. Ringe, L. Morawietz, A. Wachtel, M. Sittinger and C. Kaps (2009). "Antirheumatic drug response signatures in human chondrocytes: potential molecular targets to stimulate cartilage regeneration." Arthritis Research & Therapy **11**(1): R15.

Andreas, K., C. Lübke, T. Häupl, T. Dehne, L. Morawietz, J. Ringe, C. Kaps and M. Sittinger (2008). "Key regulatory molecules of cartilage destruction in rheumatoid arthritis: an in vitro study." Arthritis Research & Therapy **10**(1): R9.

Appleton, C. T., V. Pitelka, J. Henry and F. Beier (2007). "Global analyses of gene expression in early experimental osteoarthritis." Arthritis and Rheumatism **56**(6): 1854-1868.

Autio, R., S. Kilpinen, M. Saarela, O. Kallioniemi, S. Hautaniemi and J. Astola (2009). "Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations." BMC Bioinformatics **10** (Suppl 1): S24.

Barrett, T., D. Troup, S. Wilhite, P. Ledoux, C. Evangelista, I. Kim, M. Tomashevsky, K. Marshall, K. Phillippy, P. Sherman, R. Muertter, M. Holko, O. Ayanbule, A. Yefanov and A. Soboleva (2011). "NCBI GEO: archive for functional genomics data sets--10 years on." Nucleic Acids Research **39**(Database issue): D1005-D1010.

Benito, M., J. Parker, Q. Du, J. Wu, D. Xiang, C. Perou and J. S. Marron (2004). "Adjustment of systematic microarray data biases." Bioinformatics (Oxford, England) **20**(1): 105-114.

Brachvogel, B., F. Zaucke, K. Dave, E. Norris, J. Stermann, M. Dayakli, M. Koch, J. Gorman, J. Bateman and R. Wilson (2013). "Comparative Proteomic Analysis of Normal and Collagen IX Null Mouse Cartilage Reveals Altered Extracellular

Matrix Composition and Novel Components of the Collagen IX Interactome." Journal of Biological Chemistry **288**(19): 13481-13492.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo and M. Vingron (2001). "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." Nature Genetics **29**(4): 365-371.

Breger, J., D. Lyle, J. Shallcross, J. Langone and N. S. Wang (2009). "Defining critical inflammatory parameters for endotoxin impurity in manufactured alginate microcapsules." Journal of Biomedical Materials Research. Part B: Applied Biomaterials **91**(2): 755-765.

Cai, H., H.-J. Sun, Y.-H. Wang and Z. Zhang (2014). "Relationships of common polymorphisms in IL-6, IL-1A, and IL-1B genes with susceptibility to osteoarthritis: a meta-analysis." Clinical Rheumatology(Epub ahead of publication).

Carter, S., A. Eklund, B. Mecham, I. Kohane and Z. Szallasi (2005). "Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements." BMC Bioinformatics **6**(1): 107.

Chamberlain, C. S., S. H. Brounts, D. G. Sterken, K. I. Rolnick, G. S. Baer and R. Vanderby (2011). "Gene profiling of the rat medial collateral ligament during early healing using microarray analysis." Journal of Applied Physiology **111**(2): 552-565.

Chau, M., P. Forcinito, A. C. Andrade, A. Hegde, S. Ahn, J. C. Lui, J. Baron and O. Nilsson (2011). "Organization of the Indian hedgehog – parathyroid hormone-related protein system in the postnatal growth plate." Journal of Molecular Endocrinology **47**(1): 99-107.

Cheadle, C., Y. Cho-Chung, K. Becker and M. Vawter (2003a). "Application of z-score transformation to Affymetrix data." Applied Bioinformatics **2**(4): 209-217.

Cheadle, C., M. Vawter, W. Freed and K. Becker (2003b). "Analysis of microarray data using Z score transformation." The Journal of Molecular Diagnostics **5**(2): 73-81.

Chen, C., K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin and C. Liu (2011). "Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods." PloS ONE **6**(2): e17238.

Chen, L., D. Qanie, A. Jafari, H. Taipaleenmaki, C. Jensen, A.-M. Säämänen, M. L. N. Sanz, J. Laborda, B. Abdallah and M. Kassem (2011). "Delta-like 1/fetal antigen-1 (Dlk1/FA1) is a novel regulator of chondrogenic cell differentiation via inhibition of the Akt kinase-dependent pathway." The Journal of Biological Chemistry **286**(37): 32140-32149.

Consortium, R. G. S. P. (2004). "Genome sequence of the Brown Norway rat yields insights into mammalian evolution." Nature **428**(6982): 493-521.

de Hooge, A., F. van de Loo, M. Bennink, O. Arntz, P. de Hooge and W. van den Berg (2005). "Male IL-6 gene knock out mice developed more advanced osteoarthritis upon aging." Osteoarthritis and Cartilage **13**(1): 66-73.

Dembélé, D. (2013). "A Flexible Microarray Data Simulation Model." Microarrays **2**(2): 115-130.

Doss, F., J. Menard, M. Hauschild, H. J. Kreutzer, T. Mittlmeier, M. Müller-Steinhardt and B. Müller (2007). "Elevated IL-6 levels in the synovial fluid of osteoarthritis patients stem from plasma cells." Scandinavian Journal of Rheumatology **36**(2): 136-139.

Durinck, S., Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma and W. Huber (2005). "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis." Bioinformatics (Oxford, England) **21**(16): 3439-3440.

Durinck, S., P. Spellman, E. Birney and W. Huber (2009). "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt." Nature Protocols **4**(8): 1184-1191.

Eliasson, P., T. Andersson and P. Aspenberg (2012). "Influence of a single loading episode on gene expression in healing rat Achilles tendons." Journal of Applied Physiology **112**(2): 279-288.

Eliasson, P., T. Andersson, M. Hammerman and P. Aspenberg (2013). "Primary gene response to mechanical loading in healing rat Achilles tendons." Journal of Applied Physiology **114**(11): 1519-1526.

Elo, L., L. Lahti, H. Skottman, M. Kyläniemi, R. Lahesmaa and T. Aittokallio (2005). "Integrating probe-level expression changes across generations of Affymetrix arrays." Nucleic Acids Research **33**(22): e193-e193.

Evangelou, E. and J. P. A. Ioannidis (2013). "Meta-analysis methods for genome-wide association studies and beyond." Nature Reviews: Genetics **14**(6): 379-389.

Fernández-Tajes, J., A. Soto-Hermida, M. E. Vázquez-Mosquera, E. Cortés-Pereira, A. Mosquera, M. Fernández-Moreno, N. Oreiro, C. Fernández-López, J. L. Fernández, I. Rego-Pérez and F. J. Blanco (2014). "Genome-wide DNA methylation analysis of articular chondrocytes reveals a cluster of osteoarthritic patients." Annals of the Rheumatic Diseases **73**(4): 668-677.

Funari, V., A. Day, D. Krakow, Z. Cohn, Z. Chen, S. Nelson and D. Cohn (2007). "Cartilage-selective genes identified in genome-scale analysis of non-cartilage and cartilage gene expression." BMC Genomics **8**: 165.

Gaiteri, C., Y. Ding, B. French, G. C. Tseng and E. Sibille (2014). "Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders." Genes, Brain and Behavior **13**(1): 13-24.

Goldring, M. (2012). "Do mouse models reflect the diversity of osteoarthritis in humans?" Arthritis and rheumatism **64**(10): 3072-3075.

Gouze, J. N., E. Gouze, M. P. Popp, M. L. Bush, E. A. Dacanay, J. D. Kay, P. P. Levings, K. R. Patel, J. P. Saran, R. S. Watson and S. C. Ghivizzani (2006). "Exogenous glucosamine globally protects chondrocytes from the arthritogenic effects of IL-1beta." Arthritis Research & Therapy **8**(6): R173.

Grogan, S., S. Duffy, C. Pauli, J. Koziol, A. Su, D. D'Lima and M. Lotz (2013). "Zone-specific gene expression patterns in articular cartilage." Arthritis and Rheumatism **65**(2): 418-428.

Hawrylycz, M., E. Lein, A. Guillozet-Bongaarts, E. Shen, L. Ng, J. Miller, L. van de Lagemaat, K. Smith, A. Ebbert, Z. Riley, C. Abajian, C. Beckmann, A. Bernard, D. Bertagnolli, A. Boe, P. Cartagena, M. Chakravarty, M. Chapin, J. Chong, R. Dalley, B. Daly, C. Dang, S. Datta, N. Dee, T. Dolbeare, V. Faber, D. Feng, D. Fowler, J. Goldy, B. Gregor, Z. Haradon, D. Haynor, J. Hohmann, S. Horvath, R. Howard, A. Jeromin, J. Jochim, M. Kinnunen, C. Lau, E. Lazarz, C. Lee, T. Lemon, L. Li, Y. Li, J. Morris, C. Overly, P. Parker, S. Parry, M. Reding, J. Royall, J. Schulkin, P. Sequeira, C. Slaughterbeck, S. Smith, A. Sodt, S. Sunkin, B. Swanson, M. Vawter, D. Williams, P. Wohnoutka, R. Zielke, D. Geschwind, P. Hof, S. Smith, C. Koch, S. Grant and A. Jones (2012). "An anatomically comprehensive atlas of the adult human brain transcriptome." Nature **489**(7416): 391-399.

Haynes, W., R. Higdon, L. Stanberry, D. Collins and E. Kolker (2013). "Differential Expression Analysis for Pathways." PLoS Comput Biol **9**(3): e1002967.

Hwang, K., S. Kong, S. Greenberg and P. Park (2004). "Combining gene expression data from different generations of oligonucleotide arrays." BMC Bioinformatics **5**(1): 159.

Irizarry, R., B. Hobbs, F. Collin, Y. Beazer - Barclay, K. Antonellis, U. Scherf and T. Speed (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics **4**(2): 249-264.

Jelinsky, S., S. Rodeo, J. Li, L. Gulotta, J. Archambault and H. Seeherman (2011). "Regulation of gene expression in human tendinopathy." BMC Musculoskeletal Disorders **12**: 86.

Johnson, E., C. Li and A. Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods." Biostatistics **8**(1): 118-127.

Kasprzyk, A. (2011). "BioMart: driving a paradigm change in biological data management." Database **2011**(0): bar049.

Kim, J., K. Patel, H. Jung, W. Kuo and L. Machado (2011). "AnyExpress: Integrated toolkit for analysis of cross-platform gene expression data using a fast interval matching algorithm." BMC Bioinformatics **12**(1): 75.

Klinger, P., C. Beyer, A. Ekici, H.-D. Carl, G. Schett, B. Swoboda, F. Hennig and K. Gelse (2013). "The Transient Chondrocyte Phenotype in Human Osteophytic Cartilage: A Role of Pigment Epithelium-Derived Factor?" Cartilage **4**(3): 249-255.

Kuo, W., F. Liu, J. Trimarchi, C. Punzo, M. Lombardi, J. Sarang, M. Whipple, M. Maysuria, K. Serikawa, S. Lee, D. McCrann, J. Kang, J. Shearstone, J. Burke, D. Park, X. Wang, T. Rector, P. Ricciardi-Castagnoli, S. Perrin, S. Choi, R. Bumgarner, J. Kim, G. Short, M. Freeman, B. Seed, R. Jensen, G. Church, E. Hovig, C. Cepko, P. Park, L. Ohno-Machado and T.-K. Jenssen (2006). "A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies." Nature Biotechnology **24**(7): 832-840.

Langmead, B., C. Trapnell, M. Pop and S. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biology **10**(3): R25-10.

Larsson, O. and R. Sandberg (2006). "Lack of correct data format and comparability limits future integrative microarray research." Nature Biotechnology **24**(11): 1322-1323.

Lawrence, M., W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. Morgan and V. Carey (2013). "Software for Computing and Annotating Genomic Ranges." PLoS Comput Biol **9**(8): e1003118.

Leek, J., R. Scharpf, H. c. Bravo, D. Simcha, B. Langmead, E. Johnson, D. Geman, K. Baggerly and R. Irizarry (2010). "Tackling the widespread and critical impact of batch effects in high-throughput data." Nature Reviews: Genetics **11**(10): 733-739.

Leyh, M., A. Seitz, L. Dürselen, J. Schaumburger, A. Ignatius, J. Grifka and S. Grässel (2014a). "Subchondral bone influences chondrogenic differentiation and

collagen production of human bone marrow-derived mesenchymal stem cells and articular chondrocytes." Arthritis Research & Therapy **16**(5): 453.

Leyh, M., A. Seitz, L. Dürselen, H.-R. Springorum, P. Angele, A. Ignatius, J. Grifka and S. Grässel (2014b). "Osteoarthritic cartilage explants affect extracellular matrix production and composition in cocultured bone marrow-derived mesenchymal stem cells and articular chondrocytes." Stem Cell Research & Therapy **5**(3): 77.

Litherland, G., C. Dixon, R. Lakey, T. Robson, D. Jones, D. Young, T. Cawston and A. Rowan (2008). "Synergistic Collagenase Expression and Cartilage Collagenolysis Are Phosphatidylinositol 3-Kinase/Akt Signaling-dependent." Journal of Biological Chemistry **283**(21): 14221-14229.

Mecham, B., G. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D. Wetmore, T. Mariani, I. Kohane and Z. Szallasi (2004). "Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements." Nucleic Acids Research **32**(9): e74-e74.

Miller, J., C. Cai, P. Langfelder, D. Geschwind, S. Kurian, D. Salomon and S. Horvath (2011). "Strategies for aggregating gene expression data: The collapseRows R function." BMC Bioinformatics **12**(1): 322.

Miller, J., S. Horvath and D. Geschwind (2010). "Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways." Proceedings of the National Academy of Sciences of the United States of America **107**(28): 12698-12703.

Minogue, B., S. Richardson, L. Zeef, A. Freemont and J. Hoyland (2010). "Characterization of the human nucleus pulposus cell phenotype and evaluation of novel marker gene expression to define adult stem cell differentiation." Arthritis and Rheumatism **62**(12): 3695-3705.

Moreau, Y., S. Aerts, B. De Moor, B. De Strooper and M. Dabrowski (2003). "Comparison and meta-analysis of microarray data: from the bench to the computer desk." Trends in Genetics : TIG **19**(10): 570-577.

Nam, J., P. Perera, J. Liu, B. Rath, J. Deschner, R. Gassner, T. A. Butterfield and S. Agarwal (2011). "Sequential Alterations in Catabolic and Anabolic Gene Expression Parallel Pathological Changes during Progression of Monoiodoacetate-Induced Arthritis." PLoS ONE **6**(9): e24320.

Nimgaonkar, A., D. Sanoudou, A. Butte, J. Haslett, L. Kunkel, A. Beggs and I. Kohane (2003). "Reproducibility of gene expression across generations of Affymetrix microarrays." BMC Bioinformatics **4**(1): 27.

Ochi, K., A. Derfoul and R. Tuan (2006). "A predominantly articular cartilage-associated gene, SCRG1, is induced by glucocorticoid and stimulates chondrogenesis in vitro." Osteoarthritis and Cartilage **14**(1): 30-38.

Ostlund, G. and E. Sonnhammer (2014). "Avoiding pitfalls in gene (co)expression meta-analysis." Genomics **103**(1): 21-30.

Prasad, A., S. Kumar, C. Dessimoz, S. Bleuler, O. Laule, T. Hruz, W. Gruissem and P. Zimmermann (2013). "Global regulatory architecture of human, mouse and rat tissue transcriptomes." BMC Genomics **14**(1): 716.

Pritzker, K. P. (1994). "Animal models for osteoarthritis: processes, problems and prospects." Annals of the Rheumatic Diseases **53**(6): 406-420.

Ramasamy, A., A. Mondry, C. Holmes and D. Altman (2008). "Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets." PLoS Med **5**(9): e184.

Ramos, Y., W. den Hollander, J. Bovée, N. Bomer, R. van der Breggen, N. Lakenberg, C. Keurentjes, J. Goeman, E. Slagboom, R. Nelissen, S. Bos and I. Meulenbelt (2014). "Genes Involved in the Osteoarthritis Process Identified through Genome Wide Expression Analysis in Articular Cartilage; the RAAK Study." PloS ONE **9**(7).

Reynard, L. and J. Loughlin (2013). "The genetics and functional analysis of primary osteoarthritis susceptibility." Expert Reviews in Molecular Medicine **15**: e2.

Rockel, J., S. Bernier and A. Leask (2009). "Egr-1 inhibits the expression of extracellular matrix genes in chondrocytes by TNFalpha-induced MEK/ERK signalling." Arthritis Research & Therapy **11**(1): R8.

Rodriguez-Fontenla, C., M. Calaza, E. Evangelou, A. Valdes, N. Arden, F. Blanco, A. Carr, K. Chapman, P. Deloukas, M. Doherty, T. Esko, C. Garcés Aletá, J. Gomez-Reino Carnota, H. Helgadottir, A. Hofman, I. Jonsdottir, H. Kerkhof, M. Kloppenburg, A. McCaskie, E. Ntzani, W. Ollier, N. Oreiro, K. Panoutsopoulou, S. Ralston, Y. Ramos, J. Riancho, F. Rivadeneira, E. Slagboom, U. Styrkarsdottir, U. Thorsteinsdottir, G. Thorleifsson, A. Tsezou, A. Uitterlinden, G. Wallis, M. Wilkinson, G. Zhai, Y. Zhu, D. Felson, J. Ioannidis, J. Loughlin, A. Metspalu, I. Meulenbelt, K. Stefansson, J. van Meurs, E. Zeggini, T. Spector and A. Gonzalez (2014). "Assessment of osteoarthritis candidate genes in a meta-analysis of nine genome-wide association studies." Arthritis & Rheumatology **66**(4): 940-949.

Rudy, J. and F. Valafar (2011). "Empirical comparison of cross-platform normalization methods for gene expression data." BMC Bioinformatics **12**(1): 467.

Rung, J. and A. Brazma (2012). "Reuse of public genome-wide gene expression data." Nature Reviews: Genetics **14**(2): 89-99.

Russo, M. (2007). "How to Review a Meta-analysis." Gastroenterology & Hepatology **3**(8): 637-642.

Rustici, G., N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, J. Ison, M. Keays, N. Kurbatova, J. Malone, R. Mani, A. Mupo, R. Pedro Pereira, E. Pilicheva, J. Rung, A. Sharma, A. Tang, T. Ternent, A. Tikhonov, D. Welter, E. Williams, A. Brazma, H. Parkinson and U. Sarkans (2013). "ArrayExpress update--trends in database growth and links to data analysis tools." Nucleic Acids Research **41**(Database issue): D987-D990.

Ryu, J.-H., S. Yang, Y. Shin, J. Rhee, C.-H. Chun and J.-S. Chun (2011). "Interleukin-6 plays an essential role in hypoxia-inducible factor 2α-induced experimental osteoarthritic cartilage destruction in mice." Arthritis and Rheumatism **63**(9): 2732-2743.

Sanchez, C., M. Mateus, M.-P. Defresne, J.-M. Crielaard, J.-Y. Reginster and Y. Henrotin (2002). "Metabolism of human articular chondrocytes cultured in alginate beads. Longterm effects of interleukin 1beta and nonsteroidal antiinflammatory drugs." The Journal of Rheumatology **29**(4): 772-782.

Sarmah, C. K. and S. Samarasinghe (2010). "Microarray data integration: frameworks and a list of underlying issues." Current Bioinformatics **5**(4): 280-289.

Scheller, J., A. Chalaris, D. Schmidt-Arras and S. Rose-John (2011). "The pro- and anti-inflammatory properties of the cytokine interleukin-6." Biochimica et Biophysica Acta **1813**(5): 878-888.

Schibler, L., L. Gibbs, C. Benoist-Lasselin, C. Decraene, J. Martinovic, P. Loget, A.-L. Delezoide, M. Gonzales, A. Munnich, J.-P. Jais and L. Legeai-Mallet (2009). "New insight on FGFR3-related chondrodysplasias molecular physiopathology revealed by human chondrocyte gene expression profiling." PLoS ONE **4**(10).

Shabalin, A., H. Tjelmeland, C. Fan, C. Perou and A. Nobel (2008). "Merging two gene-expression studies via cross-platform normalization." Bioinformatics (Oxford, England) **24**(9): 1154-1160.

Sims, A., G. Smethurst, Y. Hey, M. Okoniewski, S. Pepper, A. Howell, C. Miller and R. Clarke (2008). "The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis." BMC Medical Genomics **1**(1): 42.

Sîrbu, A., H. Ruskin and M. Crane (2010). "Cross-platform microarray data normalisation for regulatory network inference." PloS ONE **5**(11).

Stock, M., C. Böhm, C. Scholtysek, M. Englbrecht, B. Fürnrohr, P. Klinger, K. Gelse, S. Gayetskyy, K. Engelke, U. Billmeier, S. Wirtz, W. van den Berg and G. Schett (2013). "Wnt inhibitory factor 1 deficiency uncouples cartilage and bone destruction in tumor necrosis factor α-mediated experimental arthritis." Arthritis and Rheumatism **65**(9): 2310-2322.

Taminau, J., C. Lazar, S. Meganck, Now, xe and Ann (2014). "Comparison of Merging and Meta-Analysis as Alternative Approaches for Integrative Gene Expression Analysis." ISRN Bioinformatics **2014**: Article ID 345106.

Taminau, J., S. Meganck, C. Lazar, D. Steenhoff, A. Coletta, C. Molter, R. Duque, V. de Schaetzen, D. Solis, H. Bersini and A. Nowe (2012). "Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages." BMC Bioinformatics **13**(1): 335.

Tibshirani, R., T. Hastie, B. Narasimhan and G. Chu (2002). "Diagnosis of multiple cancer types by shrunken centroids of gene expression." Proceedings of the National Academy of Sciences of the United States of America **99**(10): 6567-6572.

Tseng, G., D. Ghosh and E. Feingold (2012). "Comprehensive literature review and statistical considerations for microarray meta-analysis." Nucleic Acids Research **40**(9): 3785-3799.

Tsuchida, A., M. Beekhuizen, M. Rutgers, G. van Osch, J. Bekkers, A. Bot, B. Geurts, W. Dhert, D. Saris and L. Creemers (2012). "Interleukin-6 is elevated in synovial fluid of patients with focal cartilage defects and stimulates cartilage matrix production in an in vitro regeneration model." Arthritis Research & Therapy **14**(6): R262.

Valdes, A., T. Spector, A. Tamm, K. Kisand, S. Doherty, E. Dennison, M. Mangino, A. Tamm, I. Kerna, D. Hart, M. Wheeler, C. Cooper, R. Lories, N. Arden and M. Doherty (2010). "Genetic variation in the SMAD3 gene is associated with hip and knee osteoarthritis." Arthritis and Rheumatism **62**(8): 2347-2352.

Valdes, A. M., N. K. Arden, A. Tamm, K. Kisand, S. Doherty, E. Pola, C. Cooper, A. Tamm, K. R. Muir, I. Kerna, D. Hart, F. O'Neil, W. Zhang, T. D. Spector, R. A. Maciewicz and M. Doherty (2010). "A meta-analysis of interleukin-6 promoter polymorphisms on risk of hip and knee osteoarthritis." Osteoarthritis and Cartilage **18**(5): 699-704.

van Gool, S., J. Emons, J. Leijten, E. Decker, C. Sticht, J. van Houwelingen, J. Goeman, C. Kleijburg, S. Scherjon, N. Gretz, J. Wit, G. Rappold, J. Post and M. Karperien (2012). "Fetal Mesenchymal Stromal Cells Differentiating towards Chondrocytes Acquire a Gene Expression Profile Resembling Human Growth Plate Cartilage." PLoS ONE **7**(11): e44561.

Witte, F., J. Dokas, F. Neuendorf, S. Mundlos and S. Stricker (2009). "Comprehensive expression analysis of all Wnt genes and their major secreted antagonists during mouse limb development and cartilage differentiation." Gene Expression Patterns **9**(4): 215-223.

Wu, L., C. Bluguermann, L. Kyupelyan, B. Latour, S. Gonzalez, S. Shah, Z. Galic, S. Ge, Y. Zhu, F. Petrigliano, A. Nsair, S. Miriuka, X. Li, K. Lyons, G. Crooks, D. McAllister, B. Van Handel, J. Adams and D. Evseenko (2013). "Human developmental chondrogenesis as a basis for engineering chondrocytes from pluripotent stem cells." Stem Cell Reports **1**(6): 575-589.

Xu, L., A. Tan, R. Winslow and D. Geman (2008). "Merging microarray data from separate breast cancer studies provides a robust prognostic test." BMC Bioinformatics **9**(1): 125.

Yang, D. and K. Jones (2009). "Effect of alginate on innate immune activation of macrophages." J. Biomed. Mater. Res. **90A**(2): 411-418.

Yang, S., J. Kim, J.-H. Ryu, H. Oh, C.-H. Chun, B. J. Kim, B. H. Min and J.-S. Chun (2010). "Hypoxia-inducible factor-2alpha is a catabolic regulator of osteoarthritic cartilage destruction." Nature Medicine **16**(6): 687-693.

Yoshimi, M., O. Miyaishi, S. Nakamura, S.-I. Shirasawa, H. Kamochi, S. Miyatani, Y. Ikawa and T. Shinomura (2005). "Identification of genes preferentially expressed in articular cartilage by suppression subtractive hybridization." Journal of Medical and Dental Sciences **52**(4): 203-211.

Zhang, M., M. R. Pritchard, F. A. Middleton, J. A. Horton and T. A. Damron (2008). "Microarray analysis of perichondral and reserve growth plate zones identifies differential gene expressions and signal pathways." Bone **43**(3): 511-520.

Zhang, R., H. Fang, Y. Chen, J. Shen, H. Lu, C. Zeng, J. Ren, H. Zeng, Z. Li, S. Chen, D. Cai and Q. Zhao (2012). "Gene expression analyses of subchondral bone in early experimental osteoarthritis by microarray." PloS ONE **7**(2).

R packages

No peer-reviewed publications

Anita Lerch, Dimos Gaidatzis and Michael Stadler (2012). QuasR:Quantify and Annotate Short Reads in R. Version 1.2.2(unpublished). Accessed update 4 April 2014.

P. Aboyoun, H. Pages and M. Lawrence (). GenomicRanges: Representation and manipulation of genomic intervals. R package version 1.14.4. Accessed update 4 April 2014.

Florian Hahne, Steffen Durinck, Robert Ivanek, Arne Mueller, Steve Lianoglou and Ge Tan (). Gviz: Plotting data and annotation information along genomic coordinates. R package version 1.6.0. Accessed update 4 April 2014.

Jonatan Taminau (2013). inSilicoMerging: Collection of Merging Techniques for Gene Expression Data. R package version 1.6.0.

Appendix 5.1

Table SD5.1 Datasets used for rat cartilage and tendon meta-analysis. Description of headings (L to R): i) **Accession** – code for access of array data from either 'ArrayExpress' or 'Gene Expression Omnibus'. Hyperlinks are provided for direct access; ii) **Tissue** – source tissue for samples. This includes sub-compartments of tissues which may have been isolated in each study, for example, cells, fascicles, layers by laser-capture micro-dissection; iii) **Source** – body region from which tissue was derived. For musculoskeletal tissues this can be quite variable; **Donor details** – sex, breed and age (or weight) for animal sources where provided; iv) **Platform** – Manufacturer | Platform | Version; v) **Citation** – chapter references. Where citations are not given these were not available from data source and were not found following database searches. Not all datasets have been published; vi) **Number of Arrays** – the total number of arrays available from the study. The number used after quality control is provided in parentheses if this differs. This may not always equate to biological replicates. **Description**– synopsis of study and associated caveats.

Accession	Tissue (Source) Donor details	Platform	Citation (number)	Number of arrays (arrays used)	Description
Dataset I (Chapter 2)	Cartilage, tendon & fibroblasts. (Hip, Knee, Achilles, Tail and DFT and dermis.). F344, ♂, 12 wks	Illumina, RatRef-12 v1.0 Expression BeadChip	NA	40 (36)	Analysis of chondrocytes and tenocytes in three conditions – whole cartilage and tendon, monolayer and three-dimensional model cultures. Array no longer in production. Four arrays removed at quality control.
Dataset II (Chapter 3)	Cartilage and tendon. (Hip, Knee, Achilles and DFT). Wistar, ♂, 12wks	Affymetrix, Genechip Gene ST 2.0	NA	24	Analysis of matrix-free cells in three conditions – native, monolayer and three-dimensional model cultures.
E-MEXP-2672	Tendon (Tail). Sprague Dawley (SD), E21 – 6wks	Affymetrix GeneChip Rat Genome 230 2.0	NA	9	Transcription profile of rat tail tendon at three developmental stages - embryonic day 21, week 3 and week 6 post-natal.

E-GEOD-42295	Cartilage (Femoro-tibial joint) NA	Affymetrix GeneChip Rat Genome 230 2.0	(Appleton, Pitelka et al. 2007)	12 (11)	Transcription profile of rat experimental model of osteo-arthritis. Four conditions consisting of either surgical intervention or sham-operated and harvested at 2 or 8 weeks. n=12. One sample did not pass quality control.
E-GEOD-8077	Cartilage (Femoro-tibial joint). SD, ♂, 300-325 g	Affymetrix GeneChip Rat Genome 230 2.0	(Appleton, Pitelka et al. 2007)	15 (14)	Transcription profile of a rat experimental model of osteoarthritis. Groups consist of surgical intervention (medial meniscotomy), contralateral limb cartilage, and sham-operate cartilage. Five samples from each group. n=10
Linköping University, Sweden	Tendon (Fascicle)	Affymetrix Gene 1.0 ST Array	Eliasson, Andersson et al. (2012)	18	Acknowledgment

Linköping University, Sweden	Tendon (Fascicle)	Rat Gene ST 1.1 Affymetrix	(Eliasson, Andersson et al. 2013)	8	
E-GEOD-9537	Cartilage. (Growth plate zones). SD, ♂, 42-46 days	Affymetrix GeneChip Rat Genome 230 2.0	(Zhang, Pritchard et al. 2008)	8	Transcription profile of rat perichondral and reserve growth plates obtained by laser-capture micro-dissection. Four zones considered: perichondral, reserve, proliferative and hypertrophic. n=2 per zone using pooled RNA from three rats at either 42 or 46 days of age.
E-GEOD-47676	Ligament (Medial collateral). Wistar, ♂, NA	Affymetrix GeneChip Rat Genome 230 2.0	(Chamberlain, Brounts et al. 2011)	9 (3/condition)	Transcript profiling of early rat healing medial collateral ligament. Three time points – intact and day 3 or day 7 post injury.
E-GEOD-6119	Cartilage (Monolayer). Wistar, ♂,	Affymetrix GeneChip Rat	(Gouze, Gouze et al.	13 (11)	Gene expression profiling of chondrocytes in

	8weeks		Genome 230 2.0	2006)		monolayer under one of four conditions: passage 3 control, plus IL-1beta, plus glucosamine, or combination of IL-1beta and glucosamine. NA. Two arrays removed at quality control
E-GEOD-28958	Cartilage.	SD, ♀, 12-14 weeks	Affymetrix GeneChip Rat Gene 1.0 ST Array	(Nam, Perera et al. 2011)	12 (11)	Gene expression analysis from arthritis model induced by mono-iodoacetate injection. Three groups consisting of saline sham, and 5, 9 or 21 days post-injection. n=12. One sample removed at quality control.
E-GEOD-23432	Cartilage.	SD, ♂, 1 week	Affymetrix GeneChip Rat Genome 230 2.0	(Chau, Forcinito et al. 2011)	24	Micro-dissection of post-natal rat growth plates. n=5.
E-GEOD-14402	Cartilage.	Neonatal, 1day.	Affymetrix GeneChip Rat	(Rockel, Bernier et al.	8	Transcription profile of primary chondrocytes treated with TNF-alpha, DMSO, MEK1/2 inhibitor

		Genome 230 2.0	2009)		or TNF-alpha + MEK1/2 inhbitor. n=2 per condition
E-GEOD-30322	Subchondral bone. Femoro-tibial joint. SD, 10 week old, ♂.	Agilent Whole Rat Genome Microarray 4x44k G4131F	(Zhang, Fang et al. 2012)	30 (4)	Medial meniscotomy and medial collateral ligament transection plus sham controls in time course study. Normal cartilage controls used as part of preparatory study in this chapter.

Table SD5.2 Data sets used in meta-analysis of **human** microarrays from cartilage and tendon

Accession	Tissue (Source) Donor details	Platform	Citation	Number of arrays (arrays used)	Description
E-GEOD-43923	Cartilage. Femoro-tibial joint. Human, NA	Affymetrix GeneChip Human Genome U133 Plus 2.0	(Klinger, Beyer et al. 2013)	6 (6)	Articular and osteophytic cartilage obtained from knee joints following total knee arthroplasty. Total of 15 donors was used
E-GEOD-17368	Cartilage, Epiphyseal – extra digits. Human, NA	Affymetrix GeneChip Human Genome U133 Plus 2.0	NA	9 (9)	Passage post-natal epiphyseal cartilage up to passage 8. No indication of number of donors or replicates.
E-GEOD-6565	Cartilage. Distal femur. Human, 18-	Affymetrix GeneChip Human Genome U133	(Funari, Day et al. 2007)	5 (on this array)	Transcription profile of femoral cartilage from 18-22 week foetal cartilage. Other replicates performed on other arrays

	22wks	Plus 2.0			
E-GEOD-51812	Cartilage. Limb condensations. Human, 6-17 weeks	Affymetrix GeneChip Human Genome U133 Plus 2.0	(Wu, Bluguermann et al. 2013)	15	Identification of chondrocyte subsets during development. Six replicates relate to foetal limb chondrogenic condensations, six to total limb cells, and three to chondrocytes from the articular region at 17weeks. n=7
E-GEOD-40942	hfBMSCs, proximal tibial growth plate. Human, 22 week	Affymetrix GeneChip Human Genome U133 Plus 2.0	(van Gool, Emons et al. 2012)	10	Transcription profiling of human foetal mesenchymal stromal cells undergoing chondrogenic differentiation. n=1 – cells for differentiation study came from one foetus; there are six weekly time-points. Four further arrays reflect normal hyaline growth plate cartilage from adults.
E-MEXP-2488	Cartilage. Articular. Human	Affymetrix GeneChip Human Genome U133 Plus 2.0	(Minogue, Richardson et al. 2010)	3	Comparison of articular cartilage to nucleus pulposus. Three samples in this study refer to articular cartilage. Samples relating to vertebral nuclear pulposus cartilage have been deliberately excluded.
E-MEXP-2276	Cartilage. Growth plate.	Affymetrix GeneChip	(Schibler, Gibbs et al.	11	Transcription profiling of primary chondrocytes from normal foetal cartilage

	Human, foetal (18-25.5 wks) ♀/♂	Human Genome U133 Plus 2.0	2009)		or from individuals with thanatophoric dysplasia. n=11
E-GEOD-12860	Cartilage. Femoral condyles. Human, 37-74 years	Affymetrix GeneChip Human Genome HG-U133A	(Andreas, Häupl et al. 2009)	20	Transcription profiling of chondrocyte response to anti-rheumatic drugs. Passage two cells maintained in alginate bead culture. n=6 – same donors as E-GEOD-10024
E-GEOD-10024	Cartilage. Femoral condyles. Human, 37-74 years	Affymetrix GeneChip Human Genome HG-U133A	(Andreas, Lübke et al. 2008)	6	Gene expression profile of human alginate cultured chondrocytes treated with supernatant from synovial fibroblasts from RA patients. n=6. Three stimulated donor chondrocytes were pooled for each array.
E-GEOD-39795	Cartilage. Femoral condyles. Human, 23-46yo, ♀/♂	Affymetrix GeneChip Human Gene 1.0 ST	(Grogan, Duffy et al. 2013)	12	Expression profile analysis from cartilage zones. Superficial, middle, deep are considered. n=4

E-GEOD-43191	Cartilage. Femoral condyles. Human, 45-82yo, ♀/♂	Affymetrix GeneChip Human Gene 1.1 ST	(Fernández-Tajes, Soto-Hermida et al. 2014)	23	Transcription profile of articular chondrocytes from individuals with osteoarthritis. n=23
E-GEOD-26051	Tendon. Various. Human, 32-65yo, ♀/♂	Affymetrix GeneChip Human Genome U133 Plus 2.0	(Jelinsky, Rodeo et al. 2011)	46	Gene expression profiling of normal and lesional tendons from a variety of sites. n=46. Various tendon sites harvested from spectrum of donors. Reader is referred to sample descriptors in ArrayExpress.

Table SD5.3 Data sets used for consensus study using liver data sets

Accession	Tissue	Platform	Citation	Number of arrays (arrays used)	Description
E-GEOD-15238	Liver. Human, 9-12 weeks and 1.5-81 years	Affymetrix GeneChip Human Genome U133 Plus 2.0	NA	13 (13)	Embryonic and post-natal/adult livers. Total of 13 donors was used.
E-GEOD-17548	Liver. Human, NA	Affymetrix GeneChip Human Genome U133 Plus 2.0	NA	30 (30)	Tissue from cirrhosis and hepatocellular carcinoma. 30 donors.
E-GEOD-47972	Liver. Human, NA	Affymetrix GeneChip Human Genome	NA	18 (18)	Primary hepatocytes

		U133 Plus 2.0			
E-GEOD-23343	Liver. Human, NA	Affymetrix GeneChip Human Genome U133 Plus 2.0	NA	17(17)	Livers biopsies from control and type-II diabetes patients. N=17
E-GEOD-49541	Liver. Human, NA	Affymetrix GeneChip Human Genome U133 Plus 2.0	NA	72 (72)	Tissue from patients with non-alcoholic fatty liver disease

Appendix 5.2

R Codes

```
Code assumes investigator has R Expression Sets for each
gene expression profile, up-to-date R version and correct
packages in R library

Code applies weighted gene co-expression network analysis
to Ensembl annotated meta-matrices from rat and human and
outputs gene modules, network preservation statistics and
module:trait associations.

#########################################################
#Create annotation file to annotate probes across
species##
#########################################################
###Let's get the human orthologs of rat genes using
biomaRt

library(rat2302probe)

    probes<-rat2302ENTREZID
    mapped_probes<-mappedkeys(probes)
    xx<-as.list(probes[mapped_probes])
    affy.ID<-names(xx)

library(biomaRt)

#Define your database
##Ensembl does an independent mapping of array probe sequences
to genomes.  If there is no clear match then the probe is not
assigned to a gene.

     ensembl<-
     useMart("ensembl",dataset="rnorvegicus_gene_ensembl")
     listAttributes(ensembl)
###cannot call attributes from multiple databases
#get rat
     one<-
     getBM(attributes=c("affy_rat230_2","ensembl_gene_id","entre
     zgene","external_gene_id","description"),filters="affy_rat2
     30_2",values=affy.ID,mart=ensembl)
#get human
     two<-
     getBM(attributes=c("ensembl_gene_id","hsapiens_homolog_ense
     mbl_gene"),filters="affy_rat230_2",values=affy.ID,mart=ense
     mbl)
     annotation<-merge(one,two, by="ensembl_gene_id")
     table(duplicated(annotation$hsapiens_homolog_ensembl_gene))

#FALSE  TRUE
#12296  5472

     annotation2<-
     annotation[!duplicated(annotation$hsapiens_homolog_ensembl_
     gene),]
```

```
####Re-annotate human Entrez gene identifiers by Ensembl gene
identifiers

     ensembl<-useMart("ensembl",dataset="hsapiens_gene_ensembl")
# use U133 plus 2 identifiers

library(hgU133plus2.db)

    probes<-hgu133plus2ENTREZID
    mapped_probes<-mappedkeys(probes)
    xx<-as.list(probes[mapped_probes])
    U133.ID<-names(xx)

     hsap<-
     getBM(attributes=c("affy_hg_u133_plus_2","ensembl_gene_id","entr
     ezgene","external_gene_id"),filters="affy_hg_u133_plus_2",values
     =U133.ID,mart=ensembl)
     table(duplicated(hsap$ensembl_gene_id))
     hsap2<-hsap[!duplicated(hsap$ensembl_gene_id),]

     setwd("/Users/xxx/working_directory")
     save("annotation2","hsap2",
     file="Rat_to_Human_annotations.RData")

#load function to move columns in large matrices
############################
###Move columns around####
############################

moveMe <- function(data, tomove, where = "last", ba = NULL)
     {
     temp <- setdiff(names(data), tomove)
     x <- switch(
     where,
     first = data[c(tomove, temp)],
     last = data[c(temp, tomove)],
     before =
     {
     if (is.null(ba)) stop("must specify ba column")
     if (length(ba) > 1) stop("ba must be a single character string")
     data[append(temp, values = tomove, after = (match(ba, temp)-1))]
     }
     ,after =
     {
     if (is.null(ba)) stop("must specify ba column")
     if  (length(ba)  >  1)  stop("ba  must  be  a  single  character
     string")data[append(temp,  values  =  tomove,  after  =  (match(ba,
     temp)))]
     })
     x
     }

#########################################################
#Merging R expression sets from multiple data sets#####
#########################################################
#load Loess normalised Esets
#Esets prefixed with accession codes
#setwd()
```

```
library(inSilicoMerging)
     esets<-list(eset1,eset2,eset3….esetN)
     merged<-merge(esets,method="GENENORM")
detach("package:inSilicoMerging")
#rm(esets)
#plot PCA to visualize data
library(limma)
     plotMDS(exprs(merged),cex=0.5)
     boxplot(exprs(merged),las=2)

##Genenorm normalised data
     data2<-exprs(merged)
     write.csv(data2,"exprs_merged.csv")
     data2<-read.csv("exprs_merged.csv")
     colnames(data2)[1] <- c("EntrezID")

#################################
##Reannotate rat with human orthologs
#################################
setwd("/Users/xxx/working_directory")
load("Rat_to_Human_annotations.RData")

     data<-merge(data2,annotation2,by.x="EntrezID",by.y="entrezgene")
     data<-data[,-c(172:176)]
     data3<-moveMe(data,c("EntrezID"))
     data3<-moveMe(data3,c("hsapiens_homolog_ensembl_gene"),"first")
     data3<-data3[,-172]
     colnames(data3)[1]<-c("Ensembl")
     rat.data<-data3
     rm(data,data2,data3)

########################################
##Human data#############################
########################################
library(inSilicoMerging)
     esets<-list(esetHuman1, esetHuman2, esetHumanN)
     merged<-merge(esets,method="GENENORM")
detach("package:inSilicoMerging")

library(limma)
     plotMDS(exprs(merged),cex=0.5)
     data2<-exprs(merged)
     write.csv(data2,"exprs_merged.csv")
     data2<-read.csv("exprs_merged.csv")
     colnames(data2)[1] <- c("EntrezID")

########################################
##re-annotate human with Ensembl gene ids
########################################
setwd("/Users/xxx/working_directory")
load("Rat_to_Human_annotations.RData")
     data<-merge(data2,hsap2,by.x="EntrezID",by.y="entrezgene")
     data3<-moveMe(data,c("ensembl_gene_id"),"first")
     data3<-data3[,-c(123:124)]
     data3<-data3[,-2]
     colnames(data3)[1]<-c("Ensembl")
     human.data<-data3


     save(rat.data,human.data,
```

```
     file="Gnorm_cross_spp_annotated_data.RData")

#####################################################################
######Rat to human cartilage|tendon network consensus analysis###
#####################################################################
library(WGCNA)
setwd("/Users/xxx/working_directory")
load("Gnorm_cross_spp_annotated_data.RData")
#contains - rat.data - meta-matrix
#contains - human.data - meta-matrix

####Prepare Rat data
     ArrayName=names(data.frame(rat.data[,-1]))
     GeneName=rat.data$Ensembl
     exprs.rat=data.frame(t(rat.data[,-1]))
     names(exprs.rat)=rat.data[,1]
     dimnames(exprs.rat)[[1]]=names(data.frame(rat.data[,-1]))
##Prepare Human data
     ArrayName=names(data.frame(human.data[,-1]))
     GeneName=human.data$Ensembl
     exprs.human=data.frame(t(human.data[,-1]))
     names(exprs.human)=human.data[,1]
     dimnames(exprs.human)[[1]]=names(data.frame(human.data[,-1]))
####Intersect Rat data after filtering with Human data after
filtering by Ensembl ID
     commonProbesA=intersect(colnames(exprs.rat),
     colnames(exprs.human))
     length(commonProbesA)
###Remove genes not found in both datasets
     datExprRat = exprs.rat[,commonProbesA]
###intersect data on Ensembl genes
     datExprHuman = exprs.human[,commonProbesA]

##Filter each data set
###RAT####
     exprs.v=as.vector(apply(as.matrix(datExprRat),2,var,na.rm=T))
#calculate variance across the expression data
     present=as.vector(apply(!is.na(as.matrix(datExprRat)),2,sum))
     keep=exprs.v>0.4 & present>=4
     table(keep)
     filt.rat=datExprRat[,keep]

#####HUMAN
     exprs.v=as.vector(apply(
     as.matrix(datExprHuman),2,var,na.rm=T))
     present=as.vector(apply(!is.na(as.matrix(datExprHuman)),2,sum))
     keep=exprs.v>0.4 & present>=4
     table(keep)
     filt.human<-datExprHuman[,keep]

####Intersect Rat data after filtering with human data after
filtering by Ensembl ID
     commonProbesB=intersect(colnames(filt.rat),
     colnames(filt.human))
     length(commonProbesB)
#[1] 5278
##Prepare another adjacency matrix with the new data
     datExprRat=filt.rat[,commonProbesB];
##note location of comma - we are dealing with columns not rows as
the data is transposed
     datExprHuman = filt.human[,commonProbesB]
```

```
####Prepare Liver Data and annotate as per rat and human meta
matrices
load("Gnorm_LIVER_annotated_data.RData")
#humanLiver.data file created as for other human data
     ArrayName=names(data.frame(humanLIVER.data[,-1]))
     GeneName=humanLIVER.data$Ensembl
     exprs.liver=data.frame(t(humanLIVER.data[,-1]))
     names(exprs.liver)=humanLIVER.data[,1]
     dimnames(exprs.liver)[[1]]=names(data.frame(humanLIVER.data[,-
     1]))
###restrict liver data to the same genes being used for cartilage
and tendon analysis
     datExprLiver = exprs.liver[,commonProbesB]
     save(datExprHuman,datExprRat,datExprLiver,
     file="datExprRat_Human_Liver.RData")

########################
#####PICK-UP POINT 1####
########################
library(WGCNA)
setwd("/Users/xxx/working_directory")
load("datExprRat_Human_Liver.RData")

###Consensus modules ##
##transpose data again;
     t.rat<-t(datExprRat);
     t.human<-t(datExprHuman);
     options(stringsAsFactors=FALSE);
     nSets=2;
     setLabels=c("Rat","Human");
     shortLabels=c("RN","HS");

     multiExprA=vector(mode="list",length=nSets);
     multiExprA[[1]]=list(data=as.data.frame(t(t.rat)));
     names(multiExprA[[1]]$data)=colnames(datExprRat);
     colnames(multiExprA[[1]]$data)=names(datExprRat);

     multiExprA[[2]]=list(data=as.data.frame(t(t.human)));
     names(multiExprA[[2]]$data)=colnames(datExprHuman);
     colnames(multiExprA[[2]]$data)=names(datExprHuman);



##Define the Eset dimensions
     exprSize=checkSets(multiExprA)
     nGenes=exprSize$nGenes
     nSamples=exprSize$nSamples

###NAMES####
#net_A <- human and rat consensus
#multiExprA <- human and rat consensus
#consMEsA <- human and rat consensus

#net_B <- human only data
#net_C <- rat only data

#NETWORK CONSTRUCTION######
     net_A=blockwiseConsensusModules(
     multiExprA,power=7,minModuleSize=30,deepSplit=1,
     pamRespectsDendro=FALSE,mergeCutHeight=0.25,numericLabels=TRUE,
```

```
     minKMEtoStay=0,saveTOMs=FALSE,verbose=5
     )
#names(net)
     consME_A=net_A$multiMEs;
     moduleLabels=net_A$colors;
     moduleColors=labels2colors(moduleLabels)
     consTree=net_A$dendrograms[[1]];
     plotDendroAndColors(
     net_A$dendrograms[[1]],moduleColors[net_A$blockGenes[[1]]],"Cons
     ensus\nModule
     Colors",dendroLabels=FALSE,hang=0.03,addGuide=TRUE,guideHang=0.0
     5
     )

save(net_A,consMEs_A,moduleColors,consTree,file="rat_human_consensus
.RData")
rm(filt.rat,filt.human,commonProbesA,commonProbesB,t.human,t.rat,pre
sent,net,exprs.human,exprs.rat,exprs.v)
collectGarbage()

##########################
#####PICK UP POINT 2######
##########################
load("rat_to_human_consensus_net.RData")
#re-calulate the consMEs to give them colour names
     consMEsA<-multiSetMEs(multiExprA,universalColors=moduleColors);
#Add the tendon trait to the eigengenes and order them by consensus
hierarchical clustering;
     MET<-consensusOrderMEs(consMEsA);
#MET<-consensusOrderMEs(addTraitToMEs(consMEsC,tendon));
##Now call the function 'plotEigengeneNetworks' to perform the
differential analysis
     sizeGrWindow(8,10)
     par(cex=0.9)
     plotEigengeneNetworks(
     MET,setLabels,marDendro=c(0,2,2,1),
     marHeatmap=c(3,3,2,1),zlimPreservation=c(0.5,1),xLabelsAngle=90)


###LOOK AT THE PROBES IN A MODULE
#unique(colors)
     modules =c("brown");
# Select module probes
     setwd("/Users/xxx/working_directory")
     probes = names(datExprHuman)
     inModule = is.finite(match(moduleColors, modules))
     modProbes = probes[inModule]
     head(modProbes)
     module.x<-data.frame(Module=modProbes)
#module<-merge(module.x,anno,by.x="Module",by.y="EntrezID")
     write.csv(module.x,"Brown_module.csv",row.names=FALSE)

##########################
##Species-Specific Modules
##########################

##########################
#Human Specific Modules ##
##########################

##require datExprHuman - has to have consensus-matched probes
```

```
##also repeat for datExprRat
     net_B<-blockwiseModules(
     datExprHuman,power=7,minModuleSize=30,deepSplit=1,
     pamRespectsDendro = FALSE, mergeCutHeight = 0.2,
     numericLabels = TRUE, minKMEtoStay = 0, saveTOMs = FALSE,
     verbose = 5
     )

     humanMEs=net_B$MEs;
     humanLabels=net_B$colors;
     humanColors=labels2colors(humanLabels)
     humanTree=net_B$dendrograms[[1]]
     plotDendroAndColors(
     humanTree,humanColors[net_B$blockGenes[[1]]],"Module
     Colors",dendroLabels=FALSE,hang=0.03,addGuide=TRUE,guideHang=0.0
     5
     )

save(net_B,humanMEs,humanLabels,humanColors,humanTree,file="human_se
t_specific.RData")

#########################
#####PICK UP POINT 3######
#########################
load("human_set_specific.RData")
###The consensus network analysis results are represented by the
variables consMEs, moduleLabels, moduleColors, and consTree. We are
now ready to relate the human cartilage|tendon modules to the
consensus modules. We calculate the overlaps of each pair of
cartilage|tendon-consensus modules, and use the Fisher's exact test
(also known as hypergeometric test) to assign a p-value to each of
the pairwise overlaps.

#Isoloate the module labels in the order they appear in the ordered
module eigengenes
     humanModuleLabels=substring(names(humanMEs),3);
     consModuleLabels=substring(names(consME_A[[1]]$data),3)
#Convert the numeric module labels to color labels
     humanModules<-labels2colors(as.numeric(humanModuleLabels));
     consModules<-labels2colors(as.numeric(consModuleLabels));
#Numbers of affy and consensus modules
     nHumanMods=length(humanModuleLabels);
     nConsMods=length(consModuleLabels);
##initialise tables of p-values and of the corresponding counts
     pTable<-matrix(0,nrow=nHumanMods,ncol=nConsMods);
     CountTbl<-matrix(0,nrow=nHumanMods,ncol=nConsMods);
##Execute all pairwise comparisons
     for(fmod in 1:nHumanMods)
     for(cmod in 1:nConsMods)
     {
     humanMembers=(humanColors==humanModules[fmod]);
     consMembers=(moduleColors==consModules[cmod]);
     pTable[fmod,cmod]=-log10(fisher.test(humanMembers,consMembers,
     alternative="greater")$p.value);
     CountTbl[fmod,cmod]=sum(
     humanColors==humanModules[fmod]&moduleColors==consModules[cmod])
     }
#To display the p-value and count tables in an informative way we
create a color-coded table of the intersection counts.  The colours
will indicate the p-value significance:
#truncate the p values smaller than 10^{-50} to 10^{-50}
```

```
#marginal counts (really module sizes)
     humanModTotals = apply(CountTbl,1,sum);
     consModTotals = apply(CountTbl,2,sum);
#Actual plotting
     sizeGrWindow(12,7)
     par(mfrow=c(1,1))
     par(cex=1.0)
     par(mar=c(10,14,2.7,1)+0.3)
#use the function labeledheatmap to produce the color-coded table
     labeledHeatmap(Matrix=pTable,
     xLabels=paste(" ",consModules),
     yLabels=paste(" ",humanModules),
     colorLabels=TRUE,
     xSymbols=paste("Cons ",consModules,": ",consModTotals,sep=""),
     ySymbols=paste("Cartilage|Tendon",humanModules,":",
     humanModTotals,sep=""),
     textMatrix=CountTbl,
     colors=blueWhiteRed(100)[50:100],
     main="Correspondence of Human Cartilage|Tendon set-specific and
     Human:Rat consensus modules",
     cex.text=1.0,cex.lab=1.0,setStdMargins=FALSE);

###LOOK AT THE PROBES IN A MODULE
#unique(colors)
     modules =c("red");
# Select module probes
     setwd("/Users/xxx/working_directory")
     probes = names(datExprHuman)
     inModule = is.finite(match(humanColors, modules))
     modProbes = probes[inModule]
     head(modProbes)
     module.x<-data.frame(Module=modProbes)
#module<-merge(module.x,anno,by.x="Module",by.y="EntrezID")
     write.csv(module.x,"module.csv",row.names=FALSE)

##########################
#Rat-Specific Modules#####
##########################
#Requires datExprRat
     net_C<-blockwiseModules(
     datExprRat,power=7,minModuleSize=30,
     deepSplit=1,pamRespectsDendro=FALSE,
     mergeCutHeight=0.2,numericLabels=TRUE,
     minKMEtoStay=0,saveTOMs=FALSE,verbose=5
     )

     RatMEs=net_C$MEs;
     ratLabels=net_C$colors;
     ratColors=labels2colors(ratLabels)
     ratTree=net_C$dendrograms[[1]]
     plotDendroAndColors(
     ratTree,ratColors[net_C$blockGenes[[1]]],
     "ModuleColors",dendroLabels=FALSE,
     hang=0.03,addGuide=TRUE,guideHang=0.05
     )

save(net_C,RatMEs,ratLabels,ratColors,ratTree,file="rat_set_specific
.RData")


##########################
```

```
#####PICK UP POINT 4######
#########################
load(file="rat_set_specific.RData")

#Isoloate the module labels in the order they appear in the ordered
module eigengenes
     ratModuleLabels=substring(names(RatMEs),3);
     consModuleLabels=substring(names(consME_A[[1]]$data),3)
#Convert the numeric module labels to color labels
     ratModules<-labels2colors(as.numeric(ratModuleLabels));
     consModules<-labels2colors(as.numeric(consModuleLabels));
#Numbers of affy and consensus modules
     nRatMods=length(ratModuleLabels);
     nConsMods=length(consModuleLabels);

##initialise tables of p-values and of the corresponding counts
     pTable<-matrix(0,nrow=nRatMods,ncol=nConsMods);
     CountTbl<-matrix(0,nrow=nRatMods,ncol=nConsMods);

##Execute all pairwise comparisons
     for(fmod in 1:nRatMods)
     for(cmod in 1:nConsMods)
     {
     ratMembers=(ratColors==ratModules[fmod]);
     consMembers=(moduleColors==consModules[cmod]);
     pTable[fmod,cmod]=-
     log10(fisher.test(ratMembers,consMembers,alternative="greater")$
     p.value);
     CountTbl[fmod,cmod]=sum(ratColors==ratModules[fmod]&
     moduleColors==consModules[cmod])
     }

     ratModTotals = apply(CountTbl,1,sum);
     consModTotals = apply(CountTbl,2,sum);
#Actual plotting
     sizeGrWindow(12,7)
     par(mfrow=c(1,1))
     par(cex=1.0)
     par(mar=c(12,16,2.7,1)+0.3)
#use the function labeledheatmap to produce the color-coded table
     labeledHeatmap(
     Matrix=pTable,
     xLabels=paste(" ",consModules),
     yLabels=paste(" ",ratModules),
     colorLabels=TRUE,
     xSymbols=paste("Cons ",consModules,": ",consModTotals,sep=""),
     ySymbols=paste("Cartilage|Tendon",
     ratModules,":",ratModTotals,sep=""),
     textMatrix=CountTbl,
     colors=blueWhiteRed(100)[50:100],
     main="Correspondence of Rat Cartilage|Tendon set-specific and
     Human:Rat consensus modules",
     cex.text=1.0,cex.lab=1.0,setStdMargins=FALSE
     );

###LOOK AT THE PROBES IN A MODULE
#unique(colors)
     modules =c("green");
# Select module probes
     setwd("/Users/xxx/working_directory")
     probes = names(datExprRat)
```

```
     inModule = is.finite(match(ratColors, modules))
     modProbes = probes[inModule]
     head(modProbes)
     module.x<-data.frame(Module=modProbes)
#module<-merge(module.x,anno,by.x="Module",by.y="EntrezID")
     write.csv(module.x,"Rat_green_module.csv",row.names=FALSE)


##multiExpr set already created: 1=Rat; 2=Human - create a new one

     multiExpr2=list(A1=list(data=datExprRat),
     A2=list(data=datExprHuman));
     multiColor=list(A1=ratColors);

########################################
####MODULE PRESERVATION#################
########################################

#This function assesses how well a module in one study is preserved
in another study
     mp=modulePreservation(
     multiExpr2,multiColor,referenceNetworks=1,
     verbose=3,networkType="unsigned",nPermutations=30,
     maxGoldModuleSize=100,maxModuleSize=400
     );
     stats = mp$preservation$Z$ref.A1$inColumnsAlsoPresentIn.A2;
     stats2=mp$preservation$log.pBonf$ref.A1$inColumnsAlsoPresentIn.A
     2;


     modulesPreserved<data.frame(
     stats$moduleSize,
     stats$Zsummary.pres,stats2$log.p.Bonfsummary.pres
     );
     rownames(modulesPreserved)<-rownames(stats);

setwd("/Users/xxx/working_directory");
write.csv(stats,"Module_Preservation_Z_scores.csv");
rm(mp,stats,stats2,multiExpr2,multiColor,modulesPreserved)



############################################################
## Module membership (kME) and its use in comparing networks
############################################################

     MEList = moduleEigengenes(datExprRat,colors=ratColors)
     MEs = MEList$eigengenes
     colorsRat=names(table(ratColors))

     geneModuleMembership1 = signedKME(datExprRat, MEs)
     colnames(geneModuleMembership1)=
     paste("PC",colorsRat,".cor",sep=""
     );
     MMPvalue1=corPvalueStudent(
     as.matrix(geneModuleMembership1),dim(datExprRat)[[2]]
     );
     colnames(MMPvalue1)=paste("PC",colorsRat,".pval",sep="");

     Gene = rownames(t(datExprRat))
     kMEtable1=cbind(Gene,Gene,ratColors)
```

```
     for (i in 1:length(colorsRat))
     kMEtable1=cbind(
     kMEtable1,geneModuleMembership1[,i],MMPvalue1[,i]
     )
     colnames(kMEtable1)=
     c("PSID","Gene","Module",sort(
     c(colnames(geneModuleMembership1),colnames(MMPvalue1))))

setwd("/Users/alanmueller/Desktop/Thesis/Chapter_4_MetaAnalysis/DATA
")
write.csv(kMEtable1,"kMEtable1.csv",row.names=FALSE)
#or
#write.csv(kMEtable1,"Rat_modules_kME.csv",row.names=FALSE)

#Now repeat for HUMAN, using the module assignments from A2 to
determine kME values.
# First calculate MEs for A2, since we haven't done that yet

     PCs2A = moduleEigengenes(
     datExprHuman, colors=ratColors
     );
     ME_2A = PCs2A$eigengenes;
     geneModuleMembership2=signedKME(datExprHuman,ME_2A);
     colnames(geneModuleMembership1)
     =paste("PC",colorsRat,".cor",sep="");
     MMPvalue2=corPvalueStudent(
     as.matrix(geneModuleMembership2),dim(datExprHuman)[[2]]
     );
     colnames(MMPvalue2)=paste("PC",colorsRat,".pval",sep="");

     kMEtable2 = cbind(Gene,Gene,ratColors);
     for (i in 1:length(colorsRat))
     kMEtable2=cbind(
     kMEtable2,geneModuleMembership2[,i],MMPvalue2[,i]
     )
     colnames(kMEtable2)=colnames(kMEtable1)
     write.csv(kMEtable2,"kMEtable_rat_human.csv",row.names=FALSE)

##Now that we have kME values for both networks, there are a few
additional ways in which we can compare the resulting networks.  The
first thing we can do is plot the kME values of each gene in A1
against the corresponding kME values of each gene in A2. Modules
with points showing a high correlation are highly preserved.

##all genes
     par(mfrow=c(4,3))
     for (c in 1:length(colorsRat))
     {
     verboseScatterplot(
     geneModuleMembership2[,c],geneModuleMembership1[,c],
     main=colorsRat[c], xlab="kME in Human",ylab="kME in Rat")
     }

##Using all genes allows one to include all positively and
negatively correlated genes, but often also includes a lot of noise

#subset of genes originally assigned to a given module
     par(mfrow=c(4,3))
     for (c in 1:length(colorsRat))
     {
     inMod = ratColors== colorsRat[c]
```

```
     verboseScatterplot(geneModuleMembership2[inMod,c],
     geneModuleMembership1[inMod,c],main=colorsRat[c],
     xlab="kME in Human",ylab="kME in Rat")
     }

##Using only in-module genes is a visual way of assessing hub gene
conservation: if these genes show between-set correlation, then the
genes in the upper right of the plot are likely to be common hub
genes between data sets. (Hub genes are genes that show significant
correlation with MEs and high within-module connectivity and will be
discussed below.)
#The second thing we can do is determine which genes are hubs in
both networks by determine which genes have extremely high kME
values in both networks.
     topGenesKME=NULL;
     for(c in 1:length(colorsRat))
     {
     kMErank1=rank(-geneModuleMembership1[,c])
     kMErank2=rank(-geneModuleMembership2[,c])
     maxKMErank=rank(apply(cbind(kMErank1,kMErank2+.00001),1,max))
     topGenesKME=cbind(topGenesKME,Gene[maxKMErank<=20])
     }
     colnames(topGenesKME)=colorsRat;
     topGenesKME;
     consensus.hubs<-as.data.frame(topGenesKME);
write.csv(consensus.hubs,file="CONSENSUS_HUBS_rat_human.csv",row.nam
es=FALSE)

##These genes represent the top 20 genes per module based on kME in
both networks.


###############################################
##PHENOTYPIC DATA ASSOCIATIONS WITH SPECIES NETWORK MODULES
###############################################
###LOAD IN THE PHENOTYPIC DATA###############
###############################################
##Prepared binary matrix of sample membership of defined phenotypic
groups - pData.csv for the species

     setwd("/Users/xxx/working_directory")
     traitData<-read.csv("pData.csv")
     Traits=vector(mode="list",length=nSets)
     for(set in 1:nSets)
     {
     setSamples=rownames(multiExprA[[set]]$data);
     traitRows=match(setSamples,traitData$Sample);
     Traits[[set]]=list(data=traitData[traitRows,-1]);
     rownames(Traits[[set]]$data)=setSamples;
     }
     collectGarbage()

#################################################################
##Let's look at the phenotypic data for the rat modules#########
#################################################################
##require: datExprRat, pData
     dim(datExprRat)
#[1]  170 5278
     traitData = read.csv("pData.csv")
     allTraits = traitData
     ratSamples<-rownames(datExprRat)
```

```
traitRows<-match(ratSamples,allTraits$Sample)
datTraits<-allTraits[traitRows,-1]
rownames(datTraits)<-ratSamples
head(datTraits)

#Recalculate MEs with color labels
nGenes = ncol(datExprRat)
nSamples = nrow(datExprRat)
MEs0 = moduleEigengenes(datExprRat,ratColors)$eigengenes
MEs = orderMEs(MEs0)
moduleTraitCor = cor(MEs, datTraits, use = "p");
moduleTraitPvalue = corPvalueStudent(moduleTraitCor, nSamples)
textMatrix=paste(signif(moduleTraitCor,2),
"\n(",signif(moduleTraitPvalue,1),")",
sep = "")
dim(textMatrix) = dim(moduleTraitCor)
dev.new()
par(mar = c(6, 8.5, 3, 3))

labeledHeatmap(Matrix = moduleTraitCor,
         xLabels = names(datTraits),
         yLabels = names(MEs),
         ySymbols = names(MEs),
         colorLabels = FALSE,
         setStdMargins=FALSE,
         cex.lab=0.8,
         colors = blueWhiteRed(50),
         textMatrix = textMatrix,
         cex.text = 0.8,
         zlim = c(-1,1),
         main = paste("Module-trait relationships"))


#####################################################################
##Let's look at the phenotypic data for the human modules#########
#####################################################################
##require: datExprHuman, pData
dim(datExprHuman)
#[1]  166 5278
traitData = read.csv("pData.csv")
allTraits = traitData
humanSamples<-rownames(datExprHuman)
traitRows<-match(humanSamples,allTraits$Sample)
datTraits<-allTraits[traitRows,-1]
rownames(datTraits)<-humanSamples
head(datTraits)

#Recalculate MEs with color labels
nGenes = ncol(datExprHuman)
nSamples = nrow(datExprHuman)

MEs0 = moduleEigengenes(datExprHuman,humanColors)$eigengenes
MEs = orderMEs(MEs0)

moduleTraitCor = cor(MEs, datTraits, use = "p");
moduleTraitPvalue = corPvalueStudent(moduleTraitCor, nSamples)

textMatrix=paste(
signif(moduleTraitCor, 2),
"\n(", signif(moduleTraitPvalue, 1),
")", sep = "")
```

```
     dim(textMatrix) = dim(moduleTraitCor)
     dev.new()
     par(mar = c(6, 8.5, 3, 3))

     labeledHeatmap(Matrix = moduleTraitCor,
               xLabels = names(datTraits),
               yLabels = names(MEs),
               ySymbols = names(MEs),
               colorLabels = FALSE,
               setStdMargins=FALSE,
               cex.lab=0.6,
               colors = blueWhiteRed(50),
               textMatrix = textMatrix,
               cex.text = 0.6,
               zlim = c(-1,1),
               main = paste("Module-trait relationships"))

##plotting the MDS might help define the groups

###Correlations with the phenotypic data are only moderate at best
r=0.5 and this may reflect the very diverse data that we see.  As
such we should consider the consensus data.


###MODULE/TRAITS FOR CONSENSUS NETWORKS

##CONSENSUS MODULES
######Relating consensus modules to external microarray sample
information

##set up variables to contain the module_trait correlations - choose
correct multiExpr for rat or human
     exprSize=checkSets(multiExprA);
     moduleTraitCor=list()
     moduleTraitPvalue=list()
     #caluculate the correlations
     for(set in 1:nSets)
     {
     moduleTraitCor[[set]]=cor(
     consME_A[[set]]$data,Traits[[set]]$data,use="p");
     moduleTraitPvalue[[set]]=corPvalueFisher(moduleTraitCor[[set]],e
     xprSize$nSamples[set]);
     }

#We now display the module-trait relationships using a color-coded
#table.  Print the correlations and the corresponding p-values, abd
#colour-code the entries by the p-value significance.

##convert the numerical labels to colors for labelling of modules in
the plot
     MEColors<-labels2colors(
     as.numeric(substring(names(consME_A[[set]]$data),3)));
     MEColorNames<-paste("ME",MEColors,sep="");
     sizeGrWindow(10,7)

#####################################################################
#Plot the module-trait relationship table for set number 1 (Rat)
#####################################################################
##
     set=1
     textMatrix=paste(signif(
```

```
    moduleTraitCor[[set]],2),
    "\n(",signif(moduleTraitPvalue[[set]],1),
    ")",sep="");
    dim(textMatrix)=dim(moduleTraitCor[[set]])
    par(mar=c(6,8.8,3,2.2))
    labeledHeatmap(Matrix=moduleTraitCor[[set]],
                   xLabels=names(Traits[[set]]$data),
                   yLabels=MEColorNames,
                   ySymbols=MEColorNames,
                   colorLabels=FALSE,
                   colors=blueWhiteRed(50),
                   textMatrix=textMatrix,
                   setStdMargins=FALSE,
                   cex.text=0.5,
                   zlim=c(-1,1),
                   main=paste(
                   "Module—traitrelation ships in",
                   setLabels[set]));

#######################################################################
#Plot the module-trait relationship table for set number 2 (Human)
#######################################################################
##
    set=2
    textMatrix=paste(signif(moduleTraitCor[[set]],2),
    "\n(",signif(moduleTraitPvalue[[set]],1),
    ")",sep="");
    dim(textMatrix)=dim(moduleTraitCor[[set]])
    par(mar=c(6,8.8,3,2.2))
    labeledHeatmap(Matrix=moduleTraitCor[[set]],
                   xLabels=names(Traits[[set]]$data),
                   yLabels=MEColorNames,
                   ySymbols=MEColorNames,
                   colorLabels=FALSE,
                   colors=blueWhiteRed(50),
                   textMatrix=textMatrix,
                   setStdMargins=FALSE,
                   cex.text=0.5,
                   zlim=c(-1,1),
                   main=paste(
                   "Module--trait relationships in",
                   setLabels[set]));


###There are several ways of forming a measure of module-trait
relationships that summarize the two sets into one measure.  We will
form a very conservative one: for each module -trait pair we take
the correlation that has the lower absolute value in the two sets if
the two correlations have the same sign, and zero relationship if
the two correlations have opposite signs:

# Initialize matrices to hold the consensus correlation and p-value
    consensusCor=matrix(NA,nrow(moduleTraitCor[[1]]),
    ncol(moduleTraitCor[[1]]));
    consensusPvalue=matrix(NA,nrow(moduleTraitCor[[1]]),
    ncol(moduleTraitCor[[1]]));
# Find consensus negative correlations
    negative = moduleTraitCor[[1]] < 0 & moduleTraitCor[[2]] < 0;
    consensusCor[negative]=pmax(moduleTraitCor[[1]][negative],
    moduleTraitCor[[2]][negative]);
    consensusPvalue[negative]=pmax(
```

```
     moduleTraitPvalue[[1]][negative],
     moduleTraitPvalue[[2]][negative]);
# Find consensus positive correlations
     positive = moduleTraitCor[[1]] > 0 & moduleTraitCor[[2]] > 0;
     consensusCor[positive]=pmin(moduleTraitCor[[1]][positive],
     moduleTraitCor[[2]][positive]);
     consensusPvalue[positive]=pmax(
     moduleTraitPvalue[[1]][positive],
     moduleTraitPvalue[[2]][positive]);

##we display the consensus module-trait relationships again using a
color-coded table:
     textMatrix=paste(
     signif(consensusCor,2),"\n(",
     signif(consensusPvalue,1),")",sep="");
     dim(textMatrix)=dim(moduleTraitCor[[set]])
     par(mar=c(6,8.8,3,2.2));
     labeledHeatmap(Matrix=consensusCor,
                    xLabels=names(Traits[[set]]$data),
                    yLabels=MEColorNames,
                    ySymbols=MEColorNames,
                    colorLabels=FALSE,
                    colors=blueWhiteRed(50),
                    textMatrix=textMatrix,
                    setStdMargins=FALSE,
                    cex.text=0.5,
                    zlim=c(-1,1),
                    main=paste(
                    "consensus module-trait relationships across\n",
                    paste(setLabels,collapse="and"))
                    )

###############################################################
#REPEAT  ANALYSIS  USING  LIVER  META-MATRIX  AGAINST  HUMAN  OR  RAT
#DATA############################################################

######Principal Component Analysis of complete merged data sets
#used datExprHuman or datExprRat
library(FactoMineR)

     res.pca<-PCA(datExprHuman,graph=FALSE,axes=c(1,2))
     PC1 <- res.pca$ind$coord[,1]
     PC2 <- res.pca$ind$coord[,2]
##Select appropriate species groupings
##Human
     condition<-c(
     rep("alginate",26),
     rep("Monolayer",9),
     rep("Foetal Normal",3),
     rep("FoetalDysplastic",8),
     rep("Tendon Non-lesional",23),
     rep("Tendon Lesional",23),
     rep("Cartilage",16),
     rep("Differentiating",6),
     rep("OA",23),rep("Cartilage",3),
     rep("OA",3),
     rep("Differentiating",15),
     rep("Foetal Normal",5),
     rep("Cartilage",3))

##Rat
```

```
     #condition<-c(
     #rep("alginate",26),
     #rep("Monolayer",9),
     #rep("Foetal Normal",3),
     #rep("Foetal Dysplastic",8),
     #rep("Normal",23),
     #rep("Perturbed",23),
     #rep("Normal",16),
     #rep("Differentiating",6),
     #rep("Perturbed",23),
     #rep("Normal",3),
     #rep("Perturbed",3),
     #rep("Differentiating",15),
     #rep("Foetal Normal",5),
     #rep("Normal",3))

     tissue<c(
     rep("chondrocytes",46),
     rep("tenocytes",46),
     rep("chondrocytes",74))


     condition<-as.data.frame(condition)
     tissue<-as.data.frame(tissue)
     PCs <- data.frame(cbind(PC1,PC2,condition,tissue))
     PCA.comp1<-res.pca$eig[1,2]
     PCA.comp2<-res.pca$eig[2,2]

library(ggplot2)
library(RColorBrewer)

     mypalette<-brewer.pal(7,"Accent")
     mypalette<c("lightsteelblue1","goldenrod3","grey27","orange1","g
     rey0","midnightblue","firebrick1")

p2<-ggplot(PCs)
     p2<p2+geom_point(
     aes(PC1,PC2,color=condition,shape=tissue),
     size=6,alpha=0.4)
     +scale_colour_manual(values=mypalette)
     +labs(list(x=sprintf("PC1(%.1f%%)",PCA.comp1),
     y=sprintf("PC2(%.1f%%)",PCA.comp2)))
     +theme_minimal(base_size=10,base_family="Helvetica")
     +theme(legend.position="bottom",text=element_text(size=12),
     plot.title=element_text(lineheight=.8))
     +ggtitle("Human Cartilage|Tendon Meta-Set\nPrincipal Component
     Analysis")
     +scale_shape_discrete(solid=T)
p2

#####################################################################
#Export network topology to Cytoscape for visualization#############
#####################################################################
##Choose topological overlap matrix (TOM) for rat or human and
replace as appropriate

     TOM.human= TOMsimilarityFromExpr(datExprHuman, power = 7);
     TOM.rat= TOMsimilarityFromExpr(datExprRat, power = 7);
# Read in the annotation file
     load("Rat_to_Human_annotations.RData")
#Select modules
```

```
     modules    =c("green");        ##create    vector    of    modules
c("green","yellow")
#Select module probes
     probes = names(datExprHuman)
     inModule = is.finite(match(humanColors, modules));
     modProbes = probes[inModule];
     modGenes=annotation2$external_gene_id[match(modProbes,
     annotation2$hsapiens_homolog_ensembl_gene)]
# Select the corresponding Topological Overlap
     modTOM = TOM.human[inModule, inModule]
     dimnames(modTOM) = list(modProbes, modProbes)

     cyt = exportNetworkToCytoscape(
     modTOM,
     edgeFile=paste("CytoscapeInput-edges",
     paste(modules,collapse="-"),".txt", sep=""),
     nodeFile = paste("CytoscapeInput-nodes-",
     paste(modules, collapse="-"), ".txt", sep=""),
     weighted = TRUE,
     threshold = 0.02,
     nodeNames = modProbes,
     altNodeNames = modGenes,
     nodeAttr = humanColors[inModule]) ##or NULL

save(TOM.rat,TOM.human,file="TOM_for_CYTOSCAPE.RData")
```

[END]

6: *De-novo* sequencing and label-free quantification of proteins from cartilage and tendon cells

Abstract

The complex anionic matrix surrounding chondrocytes and tenocytes *in vivo* poses considerable technical problems in mass spectrometry proteomic surveys of cartilage and tendon. Highly abundant proteins dominate profiles, limit the depth of coverage and, therefore, can result in critical loss of data necessary for integration with transcriptomic data. In parallel with a transcriptome profiling study a proteomic survey of matrix-depleted chondrocytes and tenocytes was undertaken using a tandem mass-spectrometry label-free quantification approach.

The study returned a depth of coverage for proteins exceeding those found in contemporary publications considering cartilage and tendon. Moderate expression correlation was found with Affymetrix transcriptomic data, including functional annotations and pathway topology analysis. Anti-correlated elements were also defined including the ROCK-inhibitor CHORDC1 revealing regulatory aspects the systems profiled.

Three-dimensional culture systems were found to express high levels of proteins associated with oxidative phosphorylation and mitochondrial function. Fibrin and alginate cultures presented 1390 overlapping proteins including chondrogenesis-associated protein osteoactivin (GPNMB). Signalling through the PI-3K/Akt and

PPAR peroxisome pathways were also predicted in alginate bead cultures from protein abundance.

Given the role of oxidative stress in cartilage pathology and oxidative phosphorylation in stem cell differentiation and self-renewal this study contributes to a wider understanding of the response to three-dimensional model culture systems of cartilage and tendon cells. Furthermore, peri-cellular matrix depletion aids the depth of coverage in both tendon and cartilage proteomic discovery studies.

6.1: Introduction

6.1.1: General approaches in mass spectrometry

Investigators face a trade-off in proteomics studies between achieving maximum depth, or coverage, of sample analysis and retaining the *in situ* complexity of the sample. This complexity may relate to the large numbers of proteins present and/or the structural components, e.g. extra-cellular matrix, that introduces additional obstacles related to sample handling. There is a loss of connectivity, or 'chain of evidence', between the original sample and the final peptide profile wherever there is more than one protein present due to sample losses and sensitivity of detection. The ability to robustly define large numbers of proteins from peptide mass fingerprints derived from sample proteolysis has been a key break-through in high-throughput proteomics (Thiede, Höhenwarter et al. 2005).

Once solubilized, protein fractions are complex mixtures within which the dynamic range of protein abundance is manifold (Bantscheff, Lemeer et al. 2012), from moles to attamoles. Complex protein mixtures cannot be analysed efficiently and some form of resolving process, or separation methodology, is required. Traditionally gel-electrophoresis was used to separate proteins by charge and/or mass; resolved protein bands could be cut out of the gel, digested and the resultant peptides analysed by mass spectrometry (MS).

In the contemporary mass spectrometry work-flow MS instruments are coupled directly to high-performance liquid chromatography (LC) columns. The input to these columns is a peptide mixture derived from digestion of the original sample by an enzyme(s), for example trypsin. Flow through the column(s) is retarded by

various fractionation methods (ion exchange, isoelectric focusing) thereby reducing the complexity of the peptide mixture before it elutes into the MS instrument. The elution of peptides over time is a key feature in ensuring comparability between samples on the same run (Deutsch, Lam et al. 2008). Once eluted into the instrument the peptides are ionized in an instrument-dependent manner: either electrospray ionisation (ESI) or matrix-assisted laser desorption ionisation (MALDI).

Tandem mass spectrometry, often referred to as *MS/MS*, now allows the high-throughput analysis and identification of peptides. Discovery surveys of the proteome are crudely referred to as 'shot-gun' proteomics indicating the often hypothesis-free nature of these profiling studies. Intact ionised peptides that are injected into the instrument have an initial precursor ion scan, which results in a peak in the MS reading. Dynamically, the instrument selects precursor peptides, these are isolated and then subjected to collision fragmentation; the numerous resultant fragment ions for each precursor peptide produce the tandem MS spectra raw data files. This output leads on to the bioinformatics techniques, such as *de novo* sequencing, used to assimilate and infer the composition of the original protein mixture.

6.1.2: Discovery projects: comprehensive proteome coverage and quantification in mass spectrometry

Comprehensive proteome coverage is essential where different experimental variables are being considered to ensure adequate overlap in the profiles, however, a judgement on the completeness of coverage is not straightforward and may require preliminary studies to define optimal conditions for maximal protein discovery, i.e. coverage saturation. This may require multiple technical, as well as

biological, replicates. In reality comprehensive projects are likely to define 50-70% of the proteome predicted from gene models (Beck, Claassen et al. 2011). Given the advances in liquid chromatography and improved sensitivity of MS instruments and data acquisition modern LC-MS systems have the capability to identify and quantify in the region of 5,000-10,000 proteins from a given sample (Bantscheff, Lemeer et al. 2012).

Whether a protein is detectable is related to a number of technical and biological variables including: sample handling in a manner compatible with MS; solubility and digestion protocols; searches of databases with accurate annotation, and any additional post-translational modifications (PTMs) that can result in mismatches or false negatives if unidentified; finally, proteins may not be transcribed or translated within a particular condition. It is pertinent to point out that in MS studies the absence of evidence for a protein's presence is not evidence for its absence (Beck, Claassen et al. 2011). Correlation between protein abundance and gene expression data is only moderate (Thiede, Höhenwarter et al. 2005) and this uncoupling between the transcriptome and proteome reflects the different processing rates of degradation and stability between the different levels of the biological hierarchy (Beck, Claassen et al. 2011, Payne 2015). Additionally, the accurate and reproducible measurement of both mRNA and protein (de Sousa Abreu, Penalva et al. 2009), mRNA sequence signatures (Vogel, Abreu et al. 2010) and time-delay components (Wang, Wang et al. 2010) all contribute to the explanation for the variation between transcriptome and proteome profiles.

There are a number of options for defining differential abundance between samples with each having critical advantages limitations. Those most commonly used involve either metabolic (e.g. SILAC) or chemical labeling (e.g. iTRAQ),

spiked standards (QconCAT) or label-free methods (e.g. spectral counts, relative intensity) (Bantscheff, Schirle et al. 2007, Bantscheff, Lemeer et al. 2012). In this study only label-free relative intensity quantification is used where the direct mass spectrometric signal of a peptide precursor ion derived from a particular protein is compared to the equivalent in other samples and conditions.

6.1.3: Key issues in cartilage and tendon proteomics

Although not explicitly stated many of the issues arising in cartilage proteomics may be extrapolated to tendon samples. In general, all studies suffer from under-representation of low abundance proteins due to the massive dynamic range in mixtures; high abundance proteins overwhelm the analysis and obscure other relevant, but less abundant components. Pre-fractionation of samples, for example dialysis with molecular weight cut-off filters, differential extraction of cellular compartments (Rockstroh, Müller et al.) and/or the depletion of highly abundant proteins, e.g. serum albumin, hyaluron, globins etc., is critical to improving the depth, or coverage, of analysis (Wilson, Whitelock et al. 2009). Enrichment of low-abundance proteins can be achieved by equalization of protein abundance in a sample with diverse hexapeptides bound to silica beads; unbound abundant proteins, which have saturated binding sites on beads, are washed away, whilst low abundance proteins are concentrated, thereby reducing the dynamic range of the samples (Boschetti and Righetti 2008, Millioni, Tolin et al. 2011).

6.1.4: Peri-cellular matrix

As defined above proteomic surveys benefit from a reduction in the complexity of the samples. In terms of reducing the complexity of samples from cartilage and tendon the peri-cellular matrix (PCM) (defined in **Chapter 1**) needs to be considered in addition to the general extra-cellular matrix. Furthermore, the peri-

and extra-cellular matrix of monolayer and model culture systems needs to be considered if matrix-reduction strategies targeted at native tissues are used. Depletion of the PCM, beyond reducing the sample complexity, would reduce the highly anionic network surrounding the cell that can impede LC resolution, ensure changes in abundance are more comparable across samples, improve sensitivity of detection for low abundance proteins and remove the requirement for choatropic agents used to improve the solubility of collagenous samples.

The peri-cellular matrix of the chondrocyte must be considered as an autonomous transducer of biochemical and biomechanical signals from out with the chondrocyte and is considered distinct from the extra-cellular or territorial matrix (Wilusz, Sanchez-Adams et al. 2014). The PCM is a narrow cloak that aggregates a number of chondrocytes into a structure termed a 'chondron', first described by Bennighoff almost a century ago (Benninghoff 1925). Unlike the ECM the PCM is directly anchored to the plasma membrane (McLane, Chang et al. 2013) and extends variably, up to 20μm, from the cell surface. The peri-cellular matrix of tenocytes has been previously discussed in this thesis (Ritty, Roth et al. 2003).

The composition of the PCM is predominantly type VI collagen, but additionally aggrecan, hyaluron, perlecan, biglycan and type IX collagen are present; these create a capsular mesh within which the chondrocytes reside. In a proteomic analysis of the PCM Zhang, *et al* (2011) defined the presence of three type VI collagen chains, transforming growth factor-beta induced protein (TGF-βI), ADAM28 and latent-transforming growth factor beta-binding protein 2 (LTBP2) (Zhang, Jin et al. 2011). As such, the PCM represents an enriched corona of large, negatively charged proteoglycans and collagen that would require to be depleted to further reduce the complexity of chondrocyte samples.

Using particle exclusion assays McLane, *et al* (2013) found that the PCM meshwork was variable in size. The high water content means that the PCM is not visible using phase-contrast microscopy, and is easily damaged by histochemical techniques (McLane, Chang et al. 2013).

The various enzymes used to digest extra- and peri-cellular matrix have different specificities; dispase, for example, does not cleave type VI collagen (Kielty, Lees et al. 1993, Lee, Poole et al. 1997) and was used by Zhang, *et al* (2011) to differentially digest cartilage and retrieve chondrons (Zhang, Jin et al. 2011). Pronase effects, in contrast, are directed at proteoglycans in the extracellular matrix exposing the collagen fibrils rendering them more susceptible to subsequent collagenase digestion (Kuettner, Pauli et al. 1982). In Kuettner, *et al* (1982) a final single-cell suspension is obtained by the inclusion of a 0.25% trypsin digest subsequent to pronase (0.1%) and collagenase (0.4%). Others have used hyaluronidase to deplete peri-cellular hyaluron (Nishida, Knudson et al. 2003), a major component of the PCM (McLane, Chang et al. 2013).

6.1.5: Study Aims

Although representing a critical interface between the cell and the extra-cellular matrix the PCM coats the cell surface in an anionic mesh, which was hypothesized to influence the complexity of the cartilage and tendon samples.

In order to establish a coherent and comprehensive understanding of the changes in cellular protein abundance during environmental transitions to monolayer and three-dimensional model cultures a discovery survey of the chondrocyte and tenocyte proteome was undertaken using primary isolated cells, passage three cells

from monolayer and cells from three-dimensional cultures. This was undertaken in parallel to a transcriptomic survey of the same conditions (**Chapter 3**).

It was proposed that by depleting extra- and peri-cellular matrix components a greater coverage of the proteome would be possible and low abundance proteins would be identified, thereby extending the current understanding of protein expression in these tissues. Furthermore, by performing the analysis in parallel with a gene expression study validation of differentially expressed elements would be possible across two high-throughput platforms and rational targets and markers could be identified using bioinformatics techniques.

This study sought to establish an optimal peri-cellular depletion method, maximize cell yield from small tissue volumes, enrich low-abundance proteins and demonstrate consistent differential protein abundance across replicated samples.

6.2: Methods

6.2.1: Preparatory studies to define optimal matrix-depletion protocol

Sample origins

Cartilage and tendon from eight week old, male Lewis rats (n=3, 289 ± 4 g) were obtained under the previously described conditions (section **2.2.1**). Cartilage samples were derived from either pooled hip and knee cartilage or pooled shoulder cartilage. Tendon samples were derived from tail tendon or pooled Achilles and deep flexor tendon samples from the fore- and hind-limb. For preparatory studies isolated cells were defined as either native, monolayer or model culture as described in section **3.2.1** and **Figure 3.2**. Monolayer, alginate and fibrin samples were prepared as before, with the exception of the addition of 20.4 mM $CaCl_2$ to reagents used to re-suspend lyophilized thrombin according to Wang, *et al* (1995) (Wang, Pins et al. 1995).

Sample handling

The preparation, handling and storage of samples was made with respect to minimizing loss of peptides, with glass and low-adsorption plastic tubes used in preference to standard plastics where possible (Kraut, Marcellin et al. 2009, Goebel-Stengel, Stengel et al. 2011)

Peri-cellular matrix digestion protocols

Tissue dissection was undertaken on sterilized glass Petri dishes. Tissue was washed in warmed PBS; cartilage was ground using glass mortar and pestle;

tendon samples were minced as before. Tissue underwent one of two digestion protocols:

a) Shoulder cartilage or tail tendon - standard 0.4% collagenase type II (Worthington, as before) for > 12 hours at 37 °C in media free of serum and phenol-red;

b) Pooled hip and knee or Achilles and DDFT - 'triple digest' protocol, consisting of 60 minute sequential digests of: 0.1% (w/v) pronase E (protease from *Streptomyces*, > 3.5 U/mg, Sigma, #P8811), 0.4% collagenase type II, 5 U/mL hyaluronidase (400-1000U/mg Type I-S, from bovine testes, Sigma, #H3506).

A third method consisting of 0.4% collagenase type II (> 12hrs) followed by 0.25% trypsin (trypsin from bovine pancreas, >10,000 BAEE units/mg, Sigma, #T1426) (90mins) was also used as a development of the first protocol. All enzymatic solutions were filter sterilized. The triple digest protocol was also applied to cell pellets from monolayer and model cultures. Each step was performed at 37 °C in a shaking incubator. Between each digest step samples were centrifuged at ~500 x *g* for eight minutes and the resultant supernatant was discarded. After the final digest samples were washed and centrifuged in warmed, sterile PBS. Cell count, particle exclusion assay (below) and trypan-blue viability assays were performed on all samples. All protocols were otherwise performed in a cell-culture flow hood to maintain sterility and reduce keratin contamination.

Protein extraction and resolution

Isolated cells were washed twice in warmed PBS, passed through a 70 μm strainer to produce a single-cell suspension, pelleted, then re-suspended in 0.5 mL PBS and

EDTA-free protease inhibitor cocktail (cOmplete ULTRA Mini, Roche) and transferred to LoBind tubes (Eppendorf AG., Germany). Protein samples were snap-frozen in liquid nitrogen and stored at -80 °C until required. For each tenocytes and chondrocyte sample 10^3 and 10^4 cells were stored respectively. Samples were stored in Eppendorf LoBind tubes and snap frozen and stored at -80 °C until required. Samples were defrosted and sonicated on ice (10 Hz, 50% power), three times for each sample with > 1 minute intervals between sonication events. Sonicated samples were passed through 0.22 μm cellulose acetate spin columns (Spin-X, CoStar, Corning) to removed cellular debris as per the manufacturer's recommendations. Total protein concentration was estimated by photometric analysis at 650nm (Multiskan™ Microplate Spectrophotometer, Thermo Scientific) using the Pierce™ 660 nm Protein Assay (Pierce, Thermo Scientific), following the manufacturer's guidelines, against a bovine serum albumin standard curve (range 2000-25 mg/mL) based on a polynomial equation raised to the third order. To concentrate low protein yields from native and model culture samples 10 μg were concentrated using an hydroxylated silica slurry (StrataClean Resin, Agilent Technologies Inc.) according to the manufacturer's guidelines and re-suspended in standardize volumes of loading buffer for gel electrophoresis.

Gel electrophoresis and staining

Concentrated samples, including silica beads, were re-suspended in 20 μL Laemmli (2% SDS) buffer with dithiothreitol (DTT). Samples were denatured by heating to 95 °C for ten minutes then loaded onto 4-12% polyacrylamide gels (NuPAGE® Novex® 4-12% Bis-Tris Protein Gels, Life Technologies) and electrophoresed under 200 V for thirty minutes in Tris/SDS buffer (pH 8.3). Molecular weight

standards were run in parallel (Novex® Sharp Pre-stained Protein Standard, as before). Gels were washed three times in ultra-pure water and fixed in a 30% ethanol|10% acetic acid solution for thirty minutes. Silver staining was undertaken using the Pierce™ Silver Staining Kit (as before).

Particle exclusion assay

To demonstrate the absence of PCM around digested cells a particle exclusion assay (McLane, Chang et al. 2013) was performed on samples from the preliminary analysis using 7 μm microspheres (Carboxyl Latex Beads, 4% w/v, Life Technologies) instead of sheep erythrocytes. Briefly, following digestion protocol cells were washed, pelleted and re-suspended in PBS. Equal volumes (10 μL) of a cell suspension and microspheres were mixed and vizualised on a haemocytometer under a light microscope. As the presence of the PCM excludes particles from abutting the cell, following digestion particles were anticipated to be in close apposition with the cell.

6.2.2: LC-MS/MS and label-free quantification

Study design and sample preparation

Study design was comparable to that described in section **3.2.1**. Native, monolayer and three-dimensional cultures (n = 24, four biological replicates per condition, two tissues) were prepared in parallel with the microarray samples; proteomic samples were derived from independent biological replicates. All samples, tissue, monolayer or model cultures all underwent digestion protocols to deplete extra- and peri-cellular matrix as described in **3.2.1** and **Figure 3.2**. After digestion cells were washed once more in PBS, counted, pelleted and re-suspended in 1 mL

sterile PBS aliquots with the addition of protease inhibitors. Samples were stored at -80 °C in Protein LoBind tubes.

Samples were thawed and held on ice for sonication. Chondrocyte cell suspensions were sonicated first in aliquots required to obtain 2.5 x10^5 cells (equivalent to 50 μg protein at 200 pg/cell). Each aliquot was mixed on a vortex mixer and sonicated 3 x 10 s at 30% amplitude (delivered from a 3 mm probe of a Sonics Vibra Cell™, Jencons Scientific Ltd, UK) with 50 s rest time between pulses. The probe was not washed between aliquots of the same sample but between samples. Sonicated samples were held on ice.

Tryptic digestion

For chondrocytes, a volume of sample equivalent to ~50 μg protein was added to 10 μL of Strataclean Resin beads (as before) and the beads mixed for 1 min using a vortex mixer. The samples were centrifuged at 2,000 rpm for 2 min and supernatant removed and retained. For the native chondrocytes 1.25 mL of sample was added in the first round of mixing, followed by consecutive binding with a further 1.25 mL of sample. The beads were washed with 1 mL of 25 mM ammonium bicarbonate (AmBic) and re-suspended in 80 mL of 25 mM AmBic. Five millilitres of 1% (w/v) Rapigest (Waters, UK) was added to the samples which were heated for 10 min at 80 °C on a heating block with intermittent mixing (400 rpm/15 s on/off). Five millilitres of 9.2 mg/mL DTT was added and after brief mixing the samples were heated at 60 °C for 10 min. For the alkylation step 5 mL of 33 mg/mL iodoacetamide was added and the samples held at room temperature in the dark for 30 min. Five millilitres of trypsin (0.2 mg/ mL in 50 mM acetic acid; Promega Gold) was added and the sample tubes placed in a rotary

mixer at 37 °C. After 2 hrs the same amount of trypsin was added to the sample tubes, which were incubated overnight.

The samples were pulse centrifuged and 1 mL of trifluoroacetic acid (TFA) added. Of the mixed digest 0.5 mL was spotted onto pH paper to confirm acidity. The samples were incubated in a heating block at 37 °C with intermittent mixing followed by centrifugation for 30 min at 7 °C. The supernatant digests were transferred to 0.5 mL low-bind tubes (as before) and centrifuged for a further 30 min. 10 mL of each sample was transferred to total recovery vials for MS analysis. An SDS-PAGE gel of the digest was run and was clear of bands.

For tenocytes, the maximum number of cells used for all samples was standardised to the native tenocyte sample with the smallest number of cells (130,000, ~26 μg protein). Samples were processed as described above for chondrocytes.

High resolution LC-MS/MS analysis

One millilitre of digest (chondrocytes – 500 ng protein equivalent, ~ 250 ng for the tenocytes) was injected on-column and chromatographed over a 2 hr gradient using a method whereby following a survey scan at 70,000 resolution the top ten most abundant peptide ions are fragmented and measured at high resolution (35,000) in the Orbitrap analyser to a mass accuracy of 0.01 Da. Each replicate, n=12 per cell type, was eluted with 30 minutes wash stages/blank runs between conditions (native, monolayer, model cultures). Chondrocyte and tenocyte samples were run on separate days.

All peptide separations were carried out using an UltiMate® 3000 Nano LC system (Dionex, Thermo Fisher Scientific). For each analysis the sample was loaded onto

a trap column (Acclaim PepMap 100 Dionex, 2 cm x 75 μm inner diameter, C_{18}, 3 μm, 100 Å pore size) at 5 μL/min with an aqueous solution containing 0.1% (v/v) TFA and 2% (v/v) acetonitrile. After 3 min, the trap column was set in-line with an analytical column (Easy-Spray PepMap® RSLC 15cm x 75μm inner diameter, C_{18}, 2μm, 100Å) (Dionex). Peptide elution was performed by applying a mixture of solvents A and B. Solvent A was high performance LC (HPLC) grade water with 0.1% (v/v) formic acid, and solvent B was HPLC grade acetonitrile 80% (v/v) with 0.1% (v/v) formic acid. Separations were performed by applying a linear gradient of 3.8% to 50% solvent B over 95 min at 300 nL/min followed by a washing step (5 min at 99% solvent B) and an equilibration step (15 min at 3.8% solvent B).

Mass spectroscopy

Mass spectrometry was undertaken using a hybrid quadripole-Orbitrap mass spectrometer (Q Exactive™, Thermo Scientific) (Michalski, Damoc et al. 2011) operated in data dependent positive (ESI^+) mode to automatically switch between full scan MS and *MS*/*MS* acquisition. Survey full scan MS spectra (m/z 300-2000) were acquired in the Orbitrap with 70,000 resolution (m/z 200) after accumulation of ions to $1x10^6$ target value based on predictive automatic gain control (AGC) values from the previous full scan. Dynamic exclusion was set to 20 s. The ten most intense multiply charged ions ($z \geq 2$) were sequentially isolated and fragmented in the octopole collision cell by higher energy collisional dissociation (HCD) with a fixed injection time of 120 ms and 35,000 resolution. Typical mass spectrometric conditions were as follows: spray voltage, 1.9 kV, no sheath or auxillary gas flow; heated capillary temperature, 250 °C; normalised HCD collision

energy 30%. The *MS/MS* ion selection threshold was set to 1 x 10^4 counts and a 2 *m/z* isolation width was set.

6.2.3: Bioinformatics

All bioinformatics analysis was undertaken using raw data files (.XML). Analysis of raw data, including *de novo* sequencing alignment, normalisation, peptide and protein identification, and label-free relative quantification was undertaken using the PEAKS software, under license (PEAKS®, version 7, Bioinformatics Solutions Inc., Waterloo, Canada).

Tandem mass spectra were used for peptide identification using database dependent and independent methods. Database dependent methods used query matching of experimental mass spectra against theoretical peptides generated from the *Rattus norvegicus* (UniProtKB, http://www.uniprot.org/taxonomy/10116) reference proteome sequence database (accessed June, 2014) containing 27,344 entries (Magrane and Consortium 2011). Both canonical and isoform sequence data was queried.

The following search parameters were used: a) peptide mass tolerance, 10 ppm; b) fragment ion mass tolerance, 0.01 Da; c) peptide charge 1+,2+ and 3+; d) trypsin was selected as the specific cleavage agent; e) one missed cleavage was permitted. Permissible amino modifications were defined as, carbamidomethylation (fixed) and oxidation of methionine (variable). Peptide sequences were also defined directly from *MS/MS* spectra by *de novo* sequencing. A database decoy-fusion method was used in PEAKS® and FDR was set at < 1%.

Protein identification, homology searching, PTM and mutations

For all protein identifications using software modules within PEAKS® – *de novo* sequence, homology searching and post-translational modification/mutation analysis – the same filtering parameters were in place. The false discovery rate was set at 0.1%, peptide score ≥ 25, protein score ≥ 20, ≥ 2 unique peptides and an average local confidence of ≥ 80% for *de novo*-only peptides.

Label-free relative quantification and differential abundance

Label-free relative quantification was undertaken using the PEAKS Q® quantification software based upon the peptide ratios derived from mass spectra peak areas. The concomitant protein ratios were calculated from the top three unique peptide peak areas for a particular protein confidently detected in multiple samples from the same group. Peptide features across multiple samples were aligned using a combinatorial model for feature matching including a high performance retention time alignment algorithm (Lin, He et al. 2013).

Label-free quantification was undertaken within tissue groups and across all conditions or pairwise combinations of conditions as for the microarray studies. For all conditions the same parameters were used. As chondrocyte and tenocyte studies were not performed in series these studies were not comparable. Feature detection was performed separately on each sample, but relative intensities of peptide features were only calculated where features were detected in multiple samples. To ensure stringency peptide features had to be present in four out of twelve samples (i.e. one condition). For a protein to be called as significantly different between native, monolayer or model cultures a -10logP-score of 20 (the weighted sum of the –log10 P-scores of the supporting peptides and equivalent to a *p*-value of 0.01) and a fold change > 1 was set. All other values were left as

defaults. Monolayer culture samples were used as the reference samples to centre the retention time alignment. The total ion current (TIC) of the samples was used to calculate the normalisation factors; this was multiplied with the area of each feature for the calculation of a sample or group ratio.

PEAKS Q® does not provide a confidence value for each $\log_2$ ratio for differentially abundant proteins. To evaluate whether there was statistical evidence for differential abundance of identified proteins between groups the $\log_2$ ratios were exported to *R* and analysed using `Limma` (Smyth 2005), as before, section **2.2.4**. The false discovery rate was determined by the Benjamini-Hochberg method (as before) and adjusted *p*-values were set at $p < 0.05$, $\log_2$ fold-change > 1.4, and B statistic (log odds ratio) > 0 set as filtering thresholds for differential abundance as used previously (see **Chapters 2** and **3**). Only proteins that passed this two-step filtering were considered to be differentially abundant between samples.

Correlation of gene expression and protein abundance

For the native to monolayer comparison Uniprot accession codes from proteins with differential expression were converted to Entrez identifiers and common gene symbols so that they were comparable with the Affymetrix gene expression data (**Chapter 3**). These protein lists were then matched to the filtered, differentially expressed genes. Duplicate entries were removed. The $\log_2$ fold-change values were plotted and the Pearson's product-moment correlation and 95% confidence interval (CI) calculated for all elements and also for those which were only positively correlated.

General bioinformatics

The use of bioinformatics tools for gene ontology functional annotation (DAVID), re-annotation (*R* packages), pathway topology analysis (SPIA), and graphical tools has all been described in **Chapters 2** and **3**. Ingenuity® Pathway Analysis predictions of upstream regulators were based upon pairwise comparisons of filtered proteins with significant differential abundance.

6.3: Results

6.3.1: Preparatory studies

Preliminary studies sought to optimise protocols maximising cell harvests and protein yield for proteomics and to reduce the complexity of the samples by matrix depletion. In initial investigations total cell numbers harvested from cartilage were higher using the triple digest protocol of sequential pronase, collagenase and hyaluronidase ($2.1\ x10^6 \pm 6.93\ x10^5$) compared to the extended collagenase digest alone ($9.7\ x10^5 \pm 4.7x10^4$), $p = 0.02$, but samples for each technique arose from different tissue locations. The mean cell harvest from collagenase-digested tendon was $5.93\ x10^5 \pm 4.2x10^4$ cells; comparison with the triple digest protocol was not possible due to the consistent finding of large cellular aggregates or 'rafts' associated with undigested fibrillar material that did not dissociate with gentle agitation.

Total protein extracts were highly variable between replicates and were not significantly different between digest protocols. The protein profiles of native cells after each digestion protocol were compared to total protein extracts for monolayer and model cultures, **Figure 6.1** and **6.2**.

To reduce the inter-sample variability in cell counts and total protein extracted the methodology was repeated and standardised for tissue source, tissue volume and cell numbers for protein extraction with respect to different matrix depletion protocols. Additionally, to reduce the presence of undigested fibrillar material and cell aggregation a 90 minute 0.25% trypsin digest was performed after 20 hour 0.4% collagenase type II digestion. Cartilage from the hip and knee from two rats

(249 ±35 g) were harvested and pooled (0.23 ± 7x10^{-3} g); tendon was pooled from DDFT and Achilles (0.14 ± 2.4 x10^{-2} g) and tissue digested using one of the two described protocols.

Total protein was extracted from cell pellets of tenocytes (10^4) and chondrocytes (10^5) derived from each digestion protocol. Silver-stained PAGE is shown in **Figure 6.3**. In general total protein yields were higher for extended collagenase and trypsin digested tissue (range: 64-69 μg/mL) compared to the shorter triple digest protocol (range: 47-55 μg/mL) on two replicates. Total cell numbers were also greater in the longer digest period.

In particle exclusion assays micro-beads were found in close apposition with cells liberated from tissue by either digestion protocol (**Figure 6.4**) indicating that in isolated cells the peri-cellular matrix was depleted rapidly. Cell viability testing with trypan blue found that < 10% of cells stained positive in either methodology.

Given that an extended digestion protocol consisting of sequential collagenase type II (0.4%) and trypsin (0.25%) resulted in improved cell harvests and was more practical for handling large numbers of samples this was elected as the most appropriate method for matrix depletion.

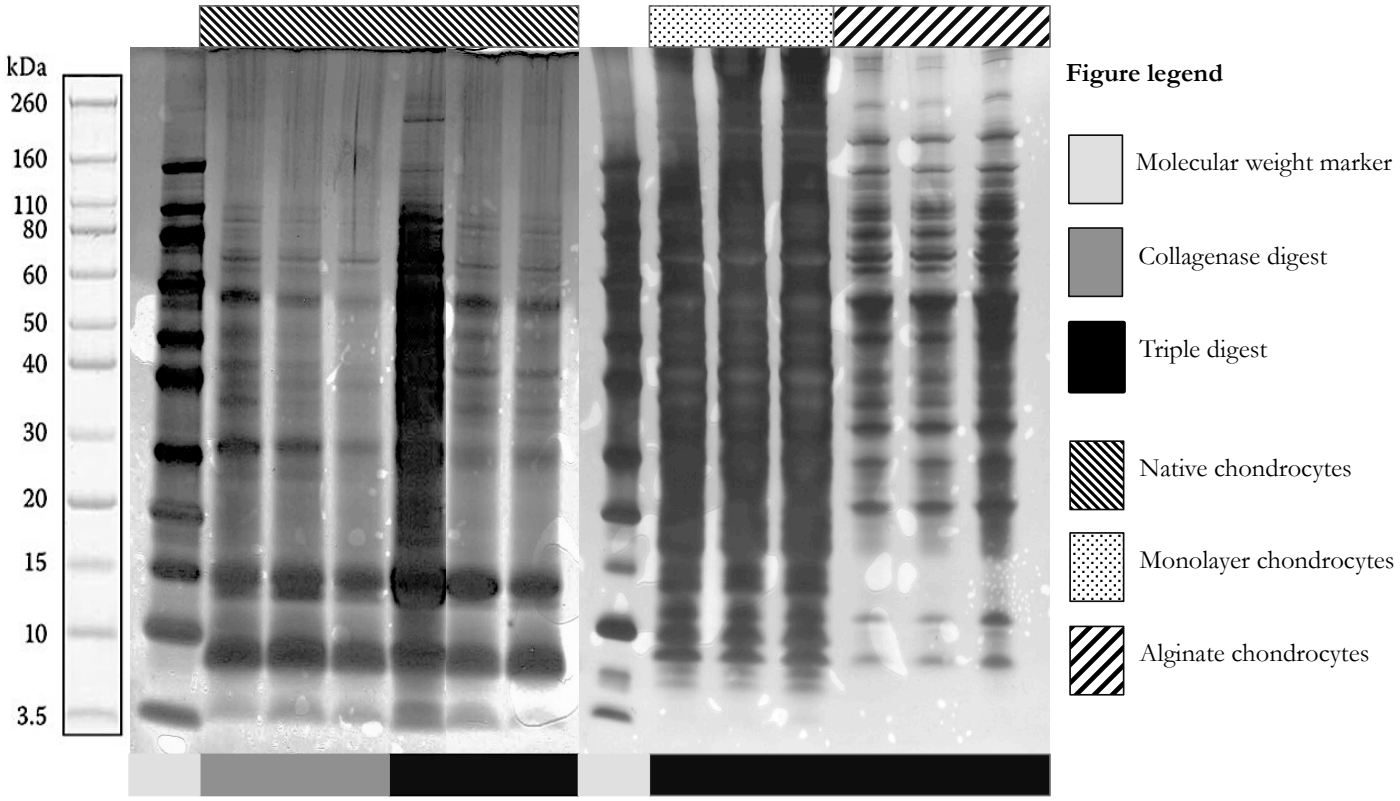


Figure 6.1:
CHONDROCYTES: Trans-illuminated, silver-stained, 4-12% polyacrylamide gel loaded with 10µg total protein resolved by electrophoresis. Images show resolved proteins from either i) native chondrocytes ii) chondrocytes from monolayer at passage three, or iii) chondrocytes from alginate following either a standard collagenase digest (native chondrocytes) or an extended serial digest protocol, all samples (**figure legend**). Native samples do not show the same complexity as samples from monolayer and alginate cultures. Differential banding is evident between conditions, however, between the two digestion protocols shown for native samples there are few differential bands except at ~30kDa where a lower MW band is found in collagenase digested chondrocytes. Within each condition there was reasonable reproducibility between biological replicates (n=3). Monolayer and alginate culture samples were run on the same gel and efficiency of the development of silver-staining accounts for the over-exposure of the monolayer samples.

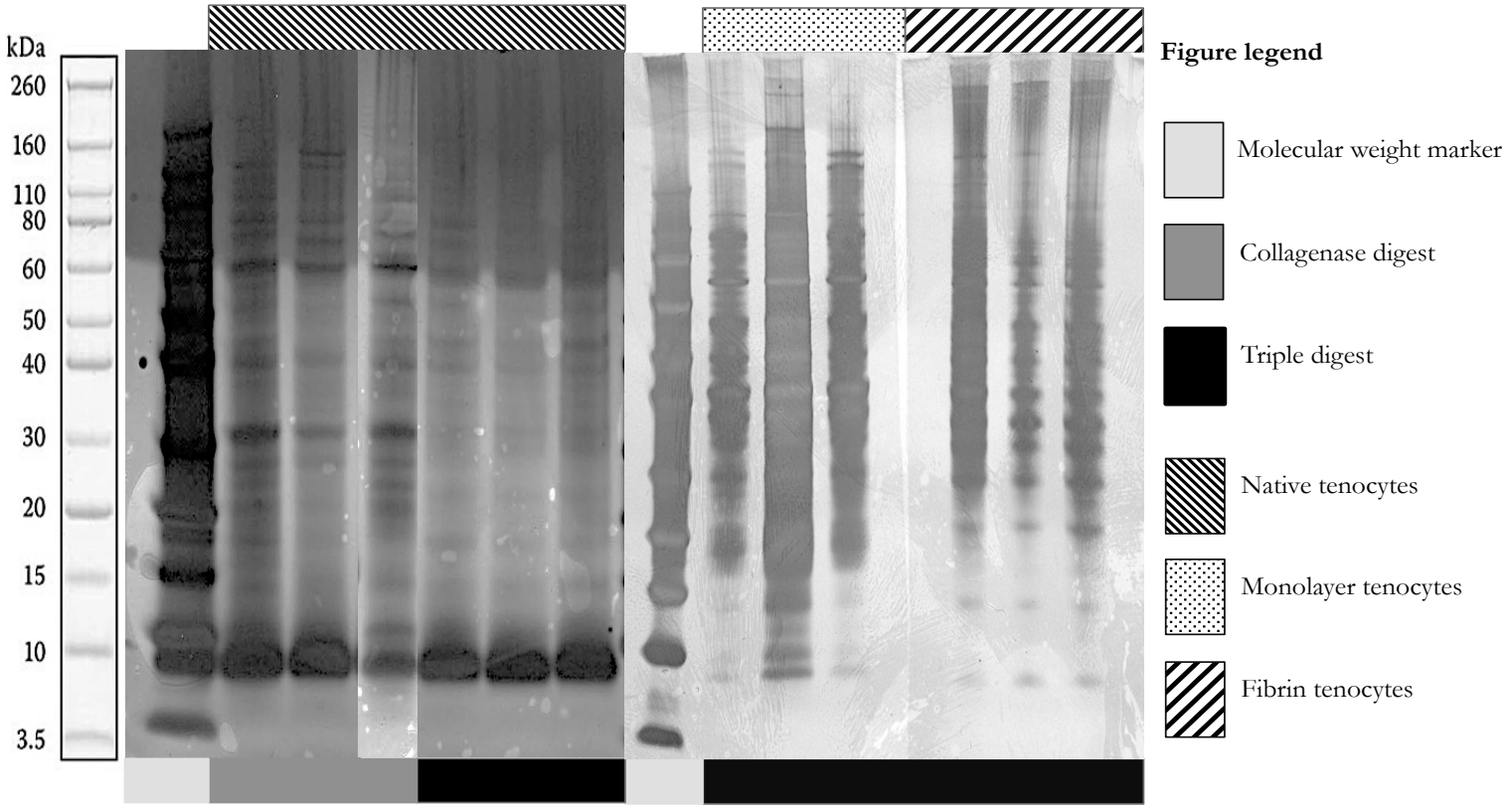


Figure 6.2:
TENOCYTES: Trans-illuminated, silver-stained, 4-12% polyacrylamide gel loaded with 10µg total protein resolved by electrophoresis. Images show resolved proteins from either i) native tenocytes ii) tenocytes from monolayer at passage three, or iii) tenocytes from fibrin following either a standard collagenase digest (native tenocytes) or an extended serial digest protocol, all samples (**figure legend**). Native samples do not show the same complexity as samples from monolayer and fibrin cultures. Differential banding is evident between conditions, however, between the two digestion protocols shown for native samples there are few differential bands other than at ~30kDa as found in **Figure 6.1**. Within each condition there was reasonable reproducibility between biological replicates (n=3).

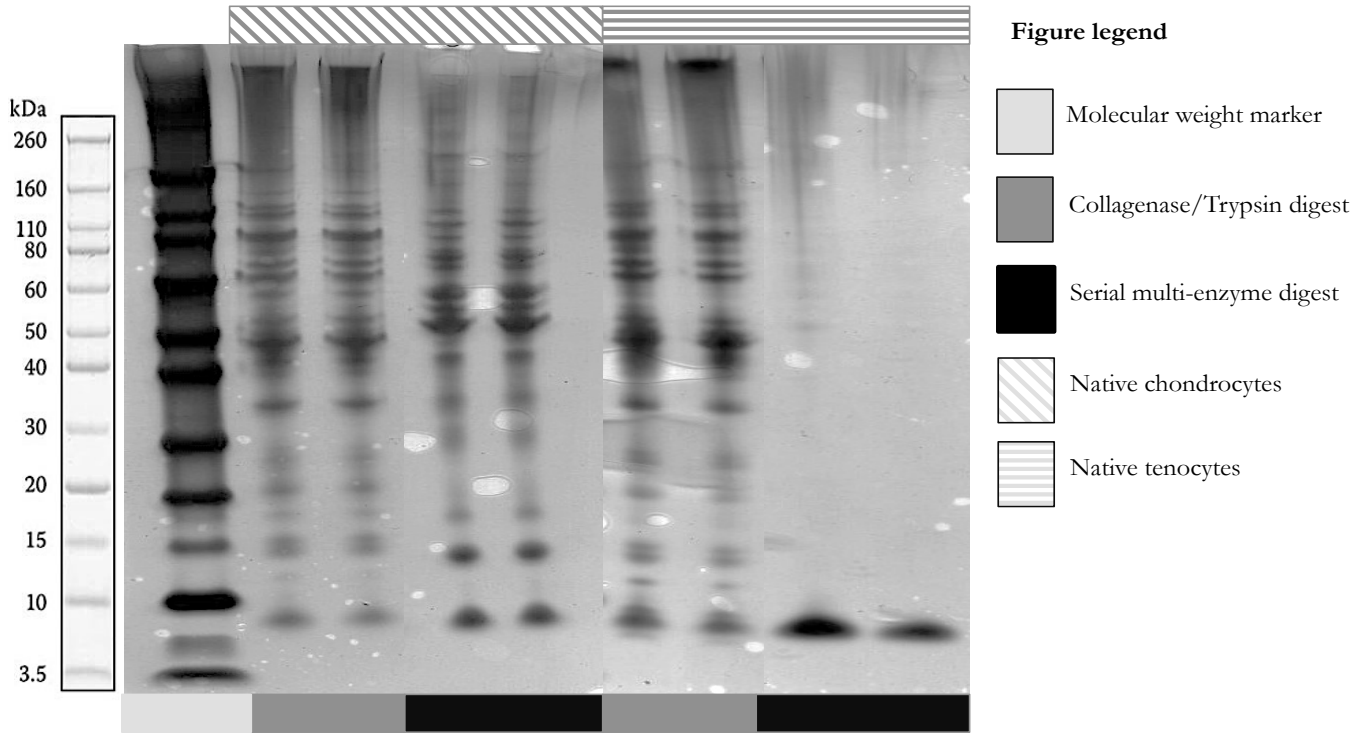
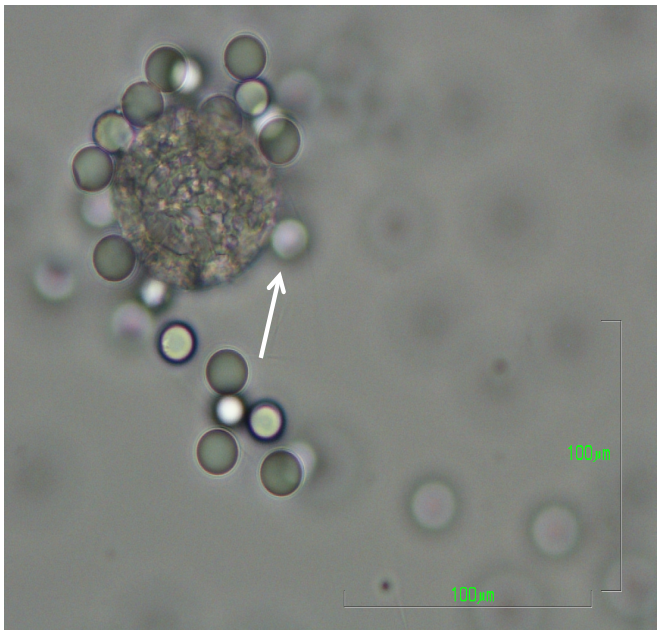
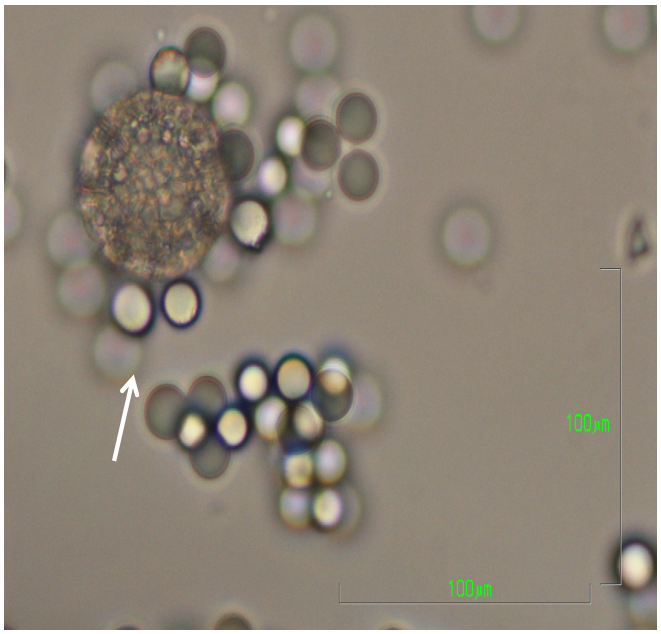


Figure 6.3:
Trans-illuminated, silver-stained, 4-12% polyacrylamide gel loaded with 10µg total protein resolved by electrophoresis. Samples represent two biological replicates. Images show resolved proteins from either native chondrocytes (10^5 cells) or tenocytes (10^4 cells) following either a 0.4% collagenase type II/0.25% trypsin sequential digest (12 hours, plus 90 minutes) or a serial digest protocol (4x90 minutes, see Methods), figure legend. Following extended digestion with collagenase and trypsin native chondrocytes and tenocytes show comparable protein bands. In contrast chondrocytes following a shorter digestion protocol using multiple enzymes displayed differential banding patterns, for example at ~15kDa. Tenocytes derived from digestion using a shorter, serial digest protocol resulted in poor total protein yields. Using standardised cell numbers rather than relying on estimated protein concentration resulted in improved reproducibility between biological samples.



A: Chrondrocyte (x40) following:
a) 0.1% pronase;
b)0.4% collagenase;
c) 50 U/mL hyaluronidase;
d)0.25% trypsin, in 90mins sequential digests



B: Chondrocyte (x40) following
a) 0.4% collagenase, 20hrs
b) 0.25% trypsin, 90 min digest

Figure 6.4: Particle exclusion assays permit the visualisation of the translucent peri-cellular matrix (PCM) which surrounds both chondrocytes and tenocytes. Where the PCM is present small particles are unable to abut the cells creating the impression of a translucent corona. Using 7 µm (scale bars) latex micro-beads the presence of the PCM after either a multi-enzyme serial digest (**A**) or a two-stage (**B**) digestion technique could not be inferred as beads (arrows) were all closely associated with isolated cells, even following the first 90 minute digest.

6.3.2: Shot-gun proteomics study with label-free quantification

Proteomics sample statistics

The wet weight of cartilage and tendon harvested from samples (n = 12) was not significantly different (p = 0.06), however, significantly greater numbers of cells were isolated from cartilage ($6.3x10^5 \pm 1.5x10^5$) relative to tendon ($3.1x10^5 \pm 1.7x10^5$), $p = 3.5x10^{-5}$. There was no statistical difference between the wet-weight of the cartilage or tendon tissue harvested for the microarray (**Chapter 3**) or proteomic samples.

Base-peak ion chromatograms

Representative base-peak ion chromatograms are presented in **Figure 6.5** (chondrocytes) and **Figure 6.6** (tenocytes) derived from separate LC runs. These demonstrated good qualitative comparisons between samples within an experimental group in terms of relative signal intensity and retention time. The chromatograms are scaled to the most abundant ions and demonstrate a large number of peaks across all samples over the whole period of analysis (two hours). A contaminant peak, possibly associated with protease inhibitors, was sharply defined at the end of each run (data not shown) and was most evident in samples where greater sample volumes were required for equivalent protein concentrations, i.e. native and model culture systems (data not shown).

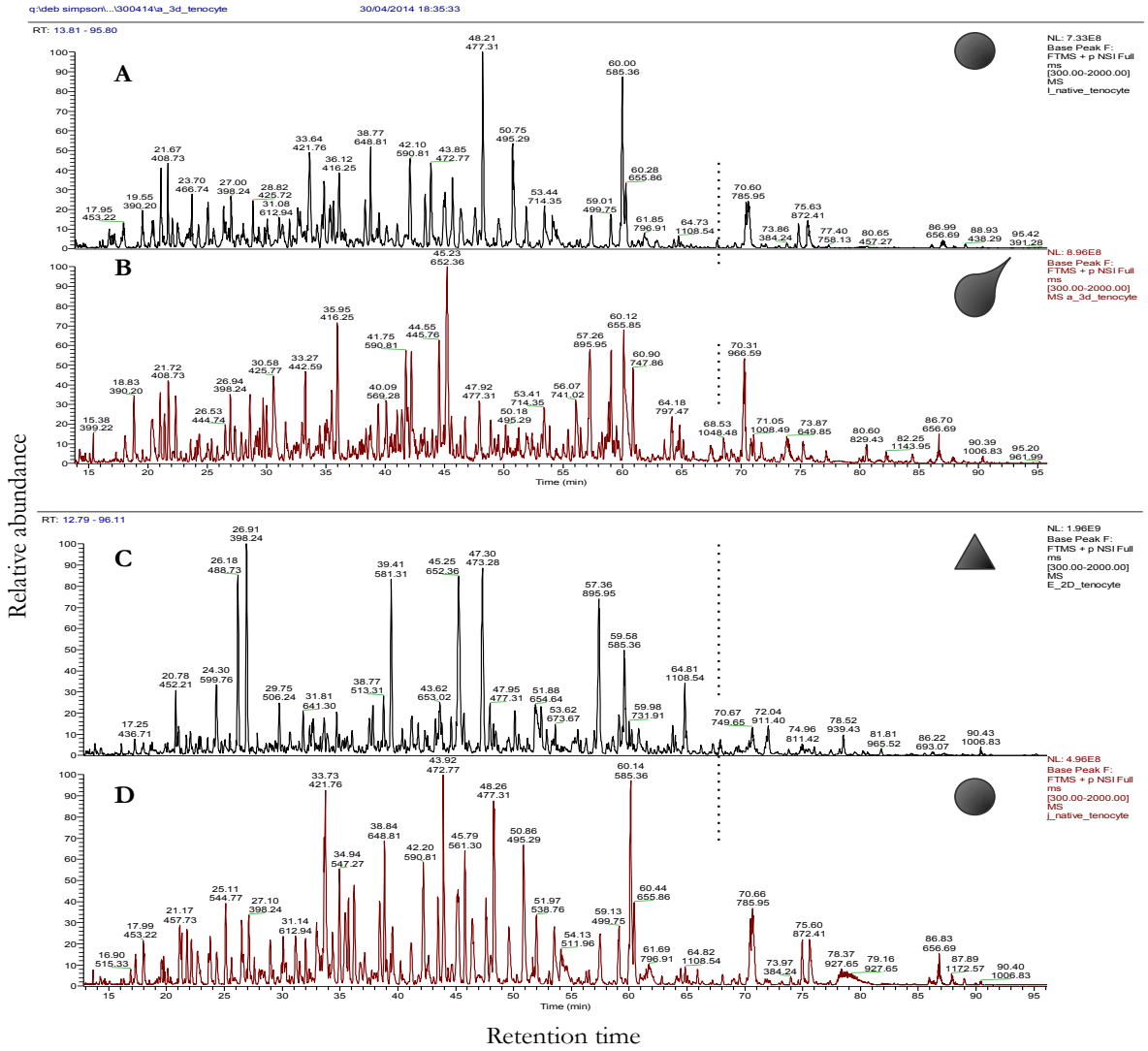


Figure 6.6: Tenocytes
Representative base-peak LC-MS ion chromatograms plotting the relative abundance/intensity of the base peak in each spectrum (y-axis) over retention time (x-axis). Peak annotations indicate the time (top value) and mass-to-charge ratio (*m/z*) (bottom value). The chromatogram is scaled to the most abundant ion.

Vertical grid lines are provided to highlight common peaks and their differences across conditions.

From top-bottom (Figure legend): **A** – Native tenocytes; **B** – fibrin tenocytes; **C** – monolayer tenocytes; **D** – native tenocytes (alternative replicate).

Legend

Native tenocytes

Monolayer tenocytes

3D culture tenocytes

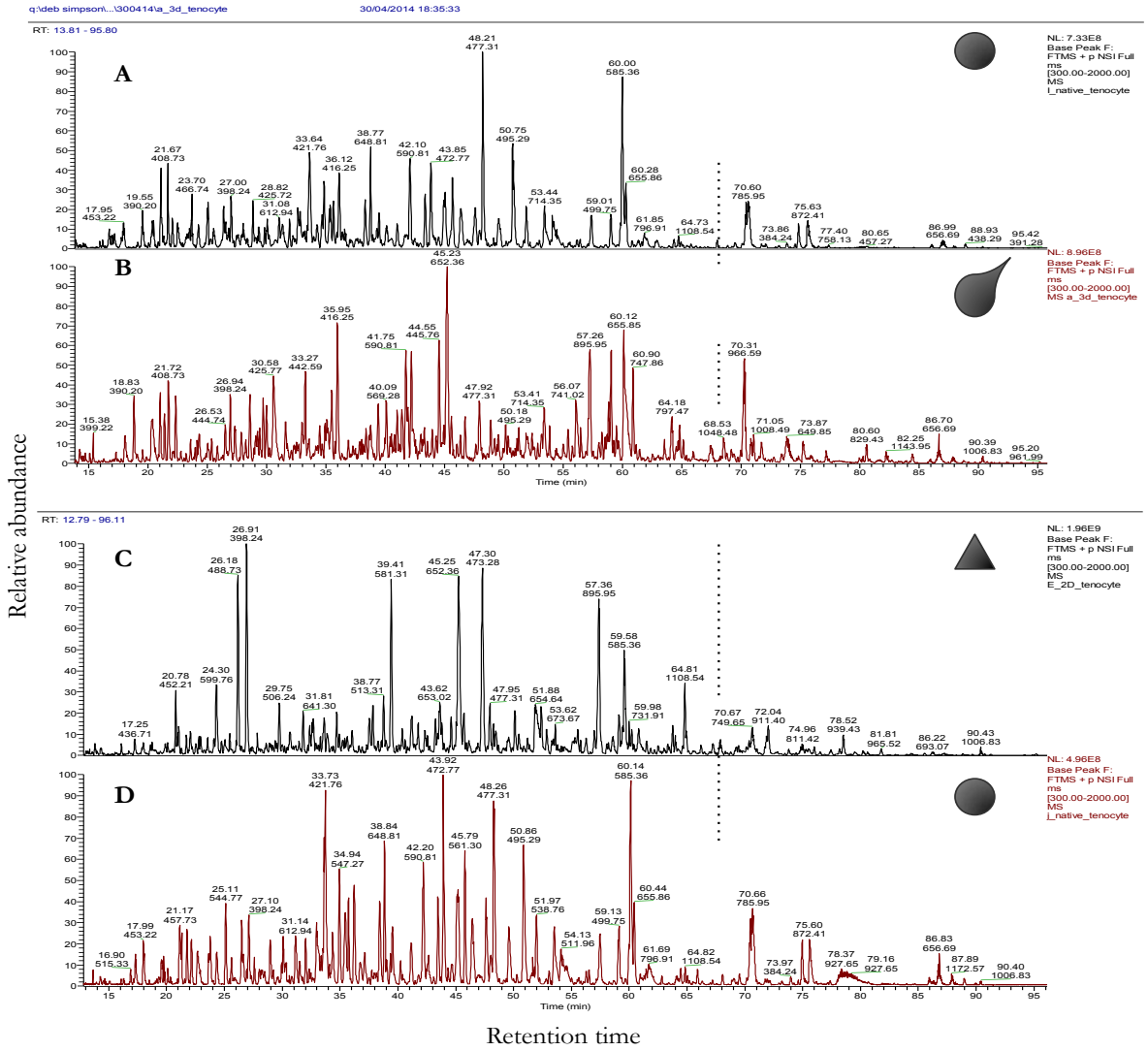


Figure 6.6: Tenocytes
Representative base-peak LC-MS ion chromatograms plotting the relative abundance/intensity of the base peak in each spectrum (y-axis) over retention time (x-axis). Peak annotations indicate the time (top value) and mass-to-charge ratio (*m/z*) (bottom value). The chromatogram is scaled to the most abundant ion.

Vertical grid lines are provided to highlight common peaks and their differences across conditions.

From top-bottom (Figure legend): **A** – Native tenocytes; **B** – fibrin tenocytes; **C** – monolayer tenocytes; **D** – native tenocytes (alternative replicate).

Legend

Native tenocytes

Monolayer tenocytes

3D culture tenocytes

MS/MS qualitative analysis and protein identification

The total number of *MS/MS* spectra and peptide spectral matches to the rat Uniprot database are presented in **Table 6.1**. In general there was considerable qualitative overlap between cells from the same condition, i.e. comparable proteins were identified. Protein groups represent collections of ambiguous (shared) peptides assigned to multiple proteins as a result of sequence similarities in the Emsembl protein database.

NATIVE CELLS

For native cells 2206 (1591 protein groups) and 1872 (1302 protein groups) were defined for chondrocytes and tenocytes respectively. This related to the identification of 540 (chondrocytes) and 237 (tenocytes) unique proteins between the native cells.

MONOLAYER

From passage three cells there were 2419 (chondrocytes) and 2154 (tenocytes) identified from 1762 and 1581 protein groups respectively. As expected there were fewer proteins unique to either chondrocytes (386) or tenocytes (191) in the monolayer condition as compared to either native cells or three-dimensional culture conditions.

MODEL CULTURES

For model cultures 2597 (1875 protein groups) and 2441 (1762 protein groups) were defined for alginate and fibrin cultures respectively. This related to 1959 (alginate) and 1839 (fibrin) unique proteins of which 1390 were confidently identified in both alginate and fibrin model cultures. Relative to fibrin cultures there were 569 unique elements found in alginate cultures; conversely 449 unique

elements were found in fibrin cultures. Complete protein identification lists are found in **SD6.**

Condition	MS Scans	MS/MS Scans	Identified Peptide Spectrum Matches	Groups (SPIDER)	Proteins (SPIDER)	*De novo* only	Without duplicate proteins	Unique to condition
Native chondrocytes	44147	93267	53453	1591 (1632)	2206 (2258)	4743	1695	540
Native tenocytes	49389	82174	43751	1302 (1326)	1872 (1904)	2973	1392	237
Monolayer chondrocytes	41212	100689	66593	1762	2419	5302	1827	386
Monolayer tenocytes	41967	98476	61254	1581	2154	5818	1632	191
Alginate model cultures	46200	89925	58430	1875	2597	3537	1959	569
Fibrin model cultures	40775	100605	62067	1762	1441	7737	1839	449

Table 6.1: Qualitative assessment of mass spectrometry data – database search parameters were the same for all analyses. Filtering settings common to all conditions: peptide FDR<0.1%; peptide -10logP score ≥25, protein -10logP score ≥20; proteins unique peptides≥2, *de novo* average local confidence ≥80%. Homology search results from SPIDER algorithm in PEAKS shown in parentheses. Unique to condition refers to a comparison of identifiers between chondrocytes and tenocytes from native, monolayer or model culture groups.

Table 6.2: Differential abundance statistics across **a**) **chondrocyte** study, **b**) **tenocyte** study. Available study proteins – total number identified across all analysis at thresholds set in Methods. Total differential abundance – remaining after statistical threshold filtering. Only unique identifiers are retained for further functional analysis. Higher/Lower – relative abundance in stated pairwise comparison.

a.

Comparison	Available study proteins	Total differential abundance	Filtered (Unique)	Higher	Lower
Native > Monolayer	1421	1031	937	158	780
Monolayer > Alginate	-	756	630	428	203
Alginate > Native	-	838	718	553	165

b.

Native > Monolayer	1402	926	835	72	763
Monolayer > Fibrin	-	884	806	384	422
Fibrin > Native	-	988	895	819	76

6.3.3: Label-free relative quantification

Dimensionality reduction

Principal component analysis of $\log_2$ ratios from either chondrocyte or tenocyte studies demonstrated strong clustering into three conditional groups: native cells, monolayer at passage three, and model cultures, **Figure 6.7**. There was strong intra-condition correlation in relative protein abundance (>0.9) for each sample.

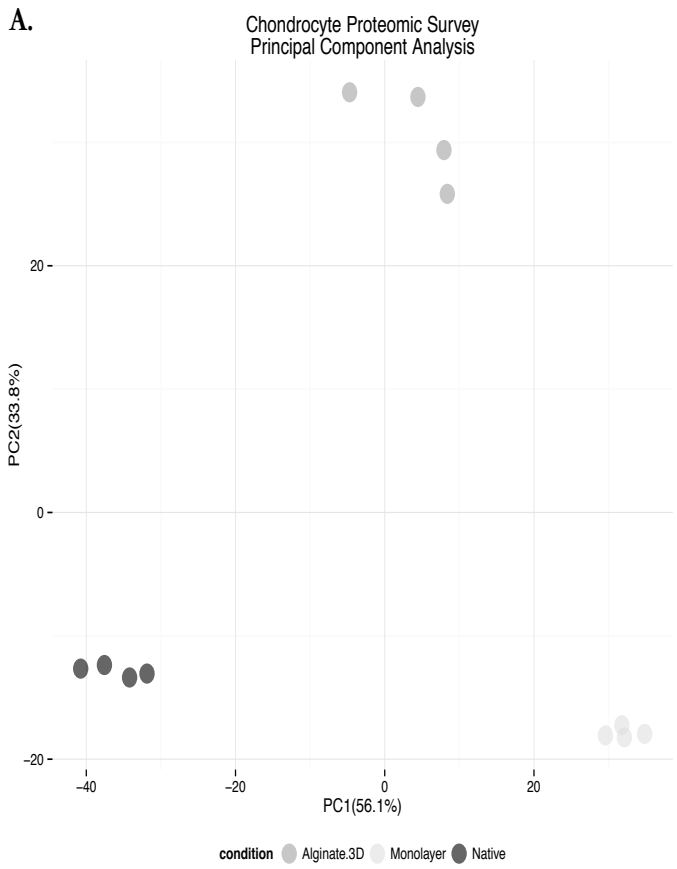
Following filtering for differential abundance, samples were assessed by unsupervised hierarchical clustering, **Figure 6.8** and **6.9**. Both chondrocyte and tenocyte studies clustered by condition. For chondrocytes there was co-clustering between native chondrocytes and alginate cultures; native tenocytes co-clustered with monolayer tenocytes with fibrin constructs within a separate clade.

Gene Ontology functional annotations

Uniprot accession codes were re-annotated with unique Entrez gene identifiers and used to functionally annotate those proteins with significant differential abundance, **Figure 6.8** and **6.9**.

Proteins more abundant in native chondrocytes were associated with biological process terms relating to 'carbohydrate metabolic process', 'regulation of biological quality' and 'positive regulation of bone resorption'.

Monolayer cells were abundant in proteins associated with 'actin filament-based process' and 'cellular metabolic process'. Three-dimensional model cultures were found to be enriched for terms associated with 'oxidation reduction' and 'generation of precursor metabolites and energy'.



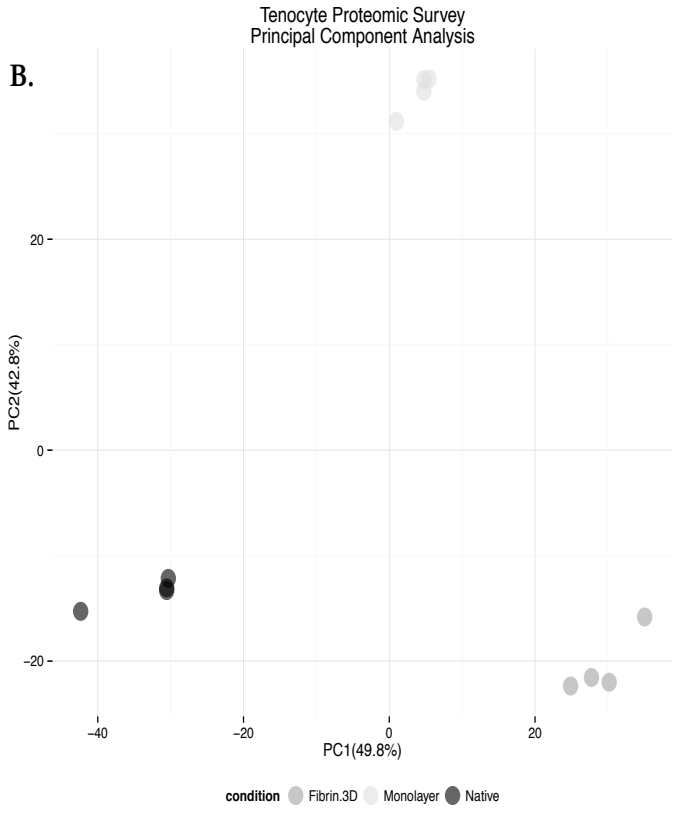
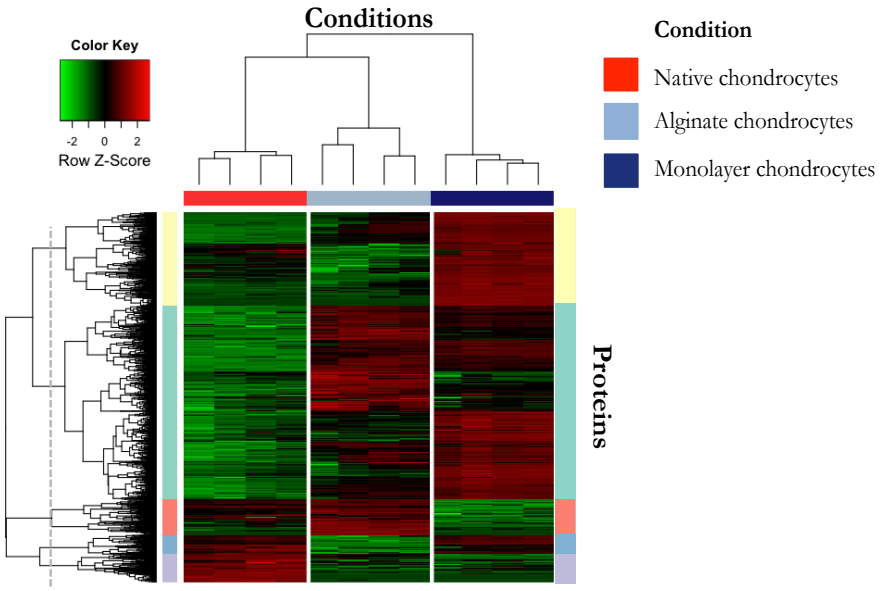


Figure 6.7: Principal Component Analysis –

A: Chondrocytes:
Plot based upon proteins with a log ratio >1 (fold-change >2) using PEAKS Q software. Correlations between samples within a condition (see figure legend) were all >0.9. Conditions show strong clustering with ~89% of the data variation described by the first two principal components.

B: Tenocytes -
Data source as for chondrocytes. First two principal components describe more than 90% of the variation in the data. As with the chondrocyte data conditions cluster robustly into 'native', 'monolayer' and 'fibrin' groups.



CC: Cytosol | Cytoplasm | Actin cytoskeleton
BP: Actin filament-based process | Protein folding | Cellular protein metabolic process
MF: Actin binding | Threonine-type peptidase activity | Translation initiation factor activity
KEGG: DNA replication

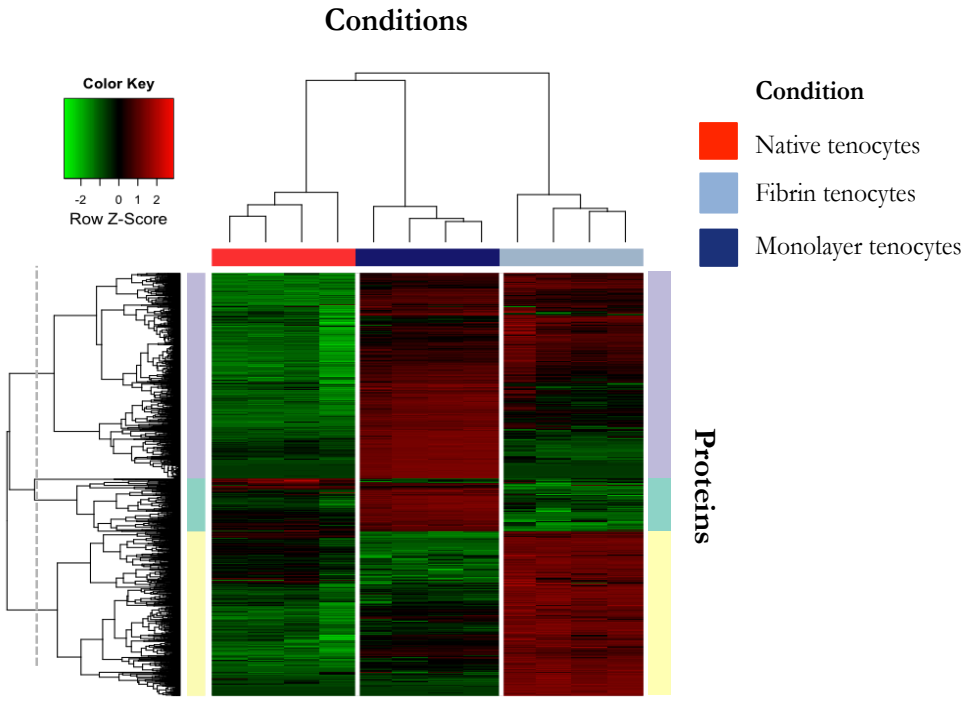
CC: Cytoplasm | Intracellular organelle part | Mitochondrial part
BP: Translation | RNA processing | Cellular metabolic process
MF: RNA binding | Nucleotide binding | Oxido-reductase activity
KEGG: Ribosome | Parkinson's disease | Oxidative phosphorylation

CC: Cytoplasm | Cell fraction | Pigment granule
BP: Establishment of localisation | Oxidation reduction | Generation of precursor metabolites and energy
MF: Oxido-reductase activity | Hydrogen ion transmembrane transporter activity | Amino-peptidase activity
KEGG: Lysosome | Oxidative phosphorylation

CC: Cytoplasm | Extracellular matrix | Intracellular part
BP: Carbohydrate metabolic process | Hexose metabolic process
MF: Pro-collagen dioxygenase activity | L-ascorbic acid binding
KEGG: No significant annotation

CC: Vesicle | Cytoplasm | Extracellular matrix
BP: Transport | Regulation of biological quality | Positive regulation of bone resorption
MF: Calcium ion binding | Heme binding | Extracellular matrix structural component conferring tensile strength
KEGG: No significant annotation

Figure 6.8 : Heatmap – chondrocyte proteins defined as having significant differential abundance by PEAKS Q software (log_2 ratios). Conditions (columns) show distinct changes in protein abundance (rows). Five groups are defined by a vertical line bisecting clades (row dendrogram) and proteins within these groups are functionally annotated using DAVID (FDR<0.01) – **CC**: Cellular component; **BP**: Biological process; **MF**: Metabolic function; **KEGG** – canonical pathway annotations. Higher abundance is show red, lower by green (**colour key**) – heatmap is scaled by row.



CC: Cytoplasm | Cytoskeleton | Melanosome
BP: Cellular process | Translation | Actin filament-based process
MF: Cytoskeletal protein binding | Translation factor activity, nucleic acid binding | Nucleotide binding
KEGG: Aminoacyl-tRNA biosynthesis | Proteosome

CC: Cytosol | Cytoplasm | Intracellular part
BP: Monosaccharide metabolic process | Organic acid metabolic process | Coenzyme metabolic process
MF: Actin binding | Oxidoreductase activity | Identical protein binding
KEGG: No significant annotation

CC: Organelle membrane | Cytoplasmic part | Mitochondrial membrane part
BP: Establishment of localisation | Generation of precursor metabolites and energy | Oxidation reduction
MF: Oxido-reductase activity | Protein transporter activity | Monovalent inorganic cation transmembrane activity
KEGG: Oxidative phosphorylation | Huntingdon's disease

Figure 6.9 : Heatmap – tenocyte proteins defined as having significant differential abundance by PEAKS Q software (log_2 ratios). Conditions (columns) show distinct changes in protein abundance (rows). Three groups are defined by a line bisecting clades (row dendrogram) and proteins within these groups are functionally annotated using DAVID (FDR<0.01) – **CC**: Cellular component; **BP**: Biological process; **MF**: Metabolic function; **KEGG** – canonical pathway annotations. Higher abundance is show red, lower by green (**colour key**) – heatmap is scaled by row.

Differential abundance

CHONDROCYTES:

After filtering and the removal of duplicate entries there were 937 proteins considered to be differentially abundant between native and monolayer chondrocytes, **Table 6.2a**. For native cells the most highly abundant proteins were dominated by erythrocyte-associated proteins: HBA, HBB, NOS2, LGALS5. Proteins related to extracellular matrix interactions including cartilage oligomeric matrix protein (COMP), chondroadherin (CHAD), integrin-binding sialoprotein (IBSP) and the collagens type II, alpha 1, and type XI, alpha 2, were more abundant in native chondrocytes. Proteins with known osteoarthritis associations were also more abundant: 14-3-3 epsilon (YWHAE), inositol triphosphate receptor, type 2 (ITPR2) and tartrate-resistant acid phosphatase type 5 (ACP5).

Monolayer chondrocytes at passage three were more abundant in proteins associated with actin filament-based processes, ACTN1, ITGB1 and debrin, and collagen fibril organization– annexin 2, collagen type III and V, and TGF-β2. The proteins thrombospondin 2 and 4, the mesoderm development candidate 2 (MESDC2), follistatin-like 2 (FSTL2) and CCN-family protein CCN2/connective tissue growth factor were also more abundant in monolayer culture than in native chondrocytes. Full differential abundance lists for all pairwise comparisons are provided in **SD6**.

Chondrocytes in alginate demonstrated higher relative abundance of chondrogenesis-associated proteins including the DCC (deleted in colorectal cancer) ligand netrin 1 (NTN1), transmembrane glycoprotein NMB/osteoactivin (GPNMB), and ADP-ribosyl cyclase 2 (encoded by bone marrow stromal cell antigen, BST1). Other notable proteins were associated with oxidation-reduction

(COX2, MAOA, TRAP1, SOD2) and nitrogen compound metabolic processes (CHI3L1, STAT3).

TENOCYTES:

There were 835 proteins that were found to be differentially abundant between native tenocytes and monolayer, **Table 6.2b**. Native tenocytes were represented by higher abundance of proteins associated with biological adhesion, cartilage oligomeric matrix protein, and the lysosomal integral membrane protein/LIMP2, encoded by SCARB2. Other lysosome-associated proteins, LAMP1 and LAMP2, were also more abundant. Proteins associated with lipid biosynthetic process, prostacyclin synthase (PTGIS) and the regulator of phosphatidylinositol levels phosphatidate cytidylyltransferase 2, encoded by CDS2, were found at higher levels in native tenocytes.

In monolayer tenocytes proteins involved with the regulation of cytoskeletal organization and microtubule dynamics were more abundant including the integrin-linked kinase (ILK) and stathmin 1 (STMN1). Actin filament-based processes were defined by the high abundance of debrin 1 (DBN1), integrin beta 1 (fibronectin receptor beta) and myosin, light chain 6 in passage three tenocytes. As found with chondrocytes MAPK1 and MAPK3 mitogen activated protein kinases, associated with regulation of gene expression, were more abundant in this monolayer condition.

Associated with tenocytes in fibrin culture was the higher expression the Bmp-antagonist gremlin 1 (GREM1) and osteoactivin (GPNMB) relative to native tenocytes. As with alginate cultures 'oxidative phosphorylation' was a significantly enriched functional annotation and was related to the higher abundance of PTGS1

and 2, MAOA and LEPREL1. The MSC marker THY-1/CD90, the fibronectin receptor integrin beta 1 and transforming growth factor, beta induced (TGF-βI), and catenin, beta 1 (CTNNB1) were all found at higher levels in fibrin constructs.

The overlap of differentially abundant proteins between the chondrocyte and tenocyte studies is presented in **Figure 6.10.** Tenocytes and chondrocytes shared differential abundance of 501 proteins in the comparison between native cells and monolayer. Fewer proteins were common to both cell types when the comparison between native cells and those in model cultures was made, 304 proteins. This is consistent with the divergence of alginate and fibrin cultures shown in the Affymetrix gene expression profiles in **Chapter 3**.

Pathway topology analysis

Over-representation analysis of KEGG pathways is provided in **Figure 6.8** and **6.9**. Using Entrez gene identifiers pathway topology analysis was used to define significantly perturbed pathways, **Figure 6.11** and **6.12**. For native to monolayer transitions the 'focal adhesion' and 'regulation of actin cytoskeleton' pathways were predicted to be activated for both chondrocytes and tenocytes, **SD6.16 - 6.17**. For chondrocytes the PI-3K-signalling pathway was shown to be perturbed in all comparisons with activation predicted in both native and monolayer cells.

In the monolayer to model culture transition for both cell types the 'Parkinson's disease' and 'Alzheimer's disease' pathways were predicted to be inhibited, however, there was conflicting evidence for the activation status of other pathways with 'focal adhesion' predicted to be activated in chondrocytes, whilst inhibited for tenocytes. For alginate cultures the 'HIF1 signaling' pathway was activated and

'PI-3K signaling' inhibited; in comparison for fibrin cultures the 'Jak-STAT' pathway was predicted to be activated along with 'regulation of actin cytoskeleton'.

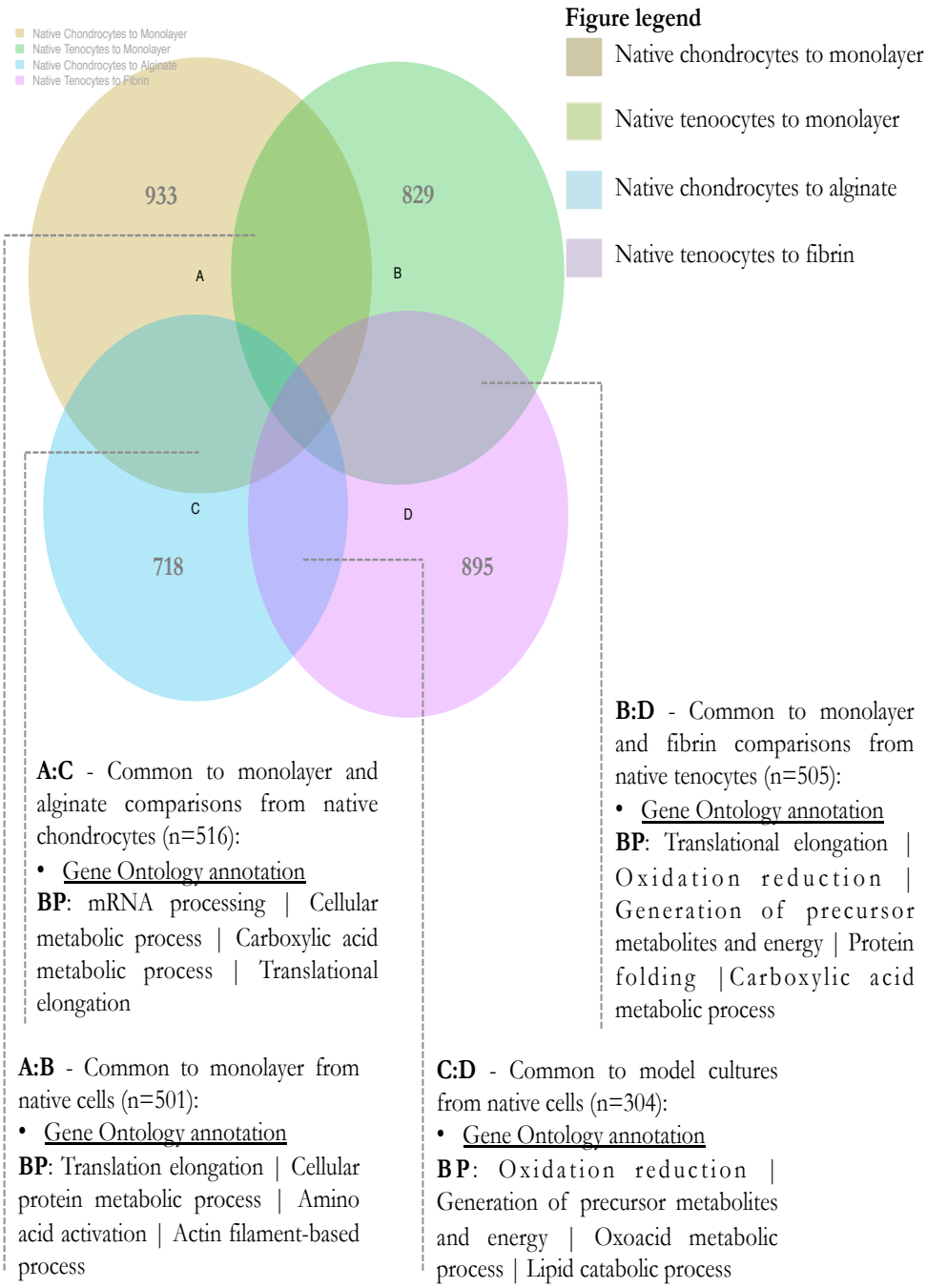


Figure 6.10 – Euler diagram showing shared proteins for each comparison across chondrocyte and tenocyte studies (**figure legend**). Common proteins are annotated by biological process terms using DAVID. **A:C** represents proteins common to both comparisons in chondrocytes; likewise, **B:D** shows proteins common to tenocyte comparisons. Across chondrocyte and tenocyte studies gene ontology studies show enrichment for biological process annotations found when studies are considered independently. There is greater overlap in proteins between native to monolayer comparisons (**A:B**) than between native to model culture comparisons (**C:D**); this concurs with the greater overlap of differentially expressed genes from monolayer and a divergence in profile for model conditions shown by the Affymetrix data in **Chapter 3**.

Figure 6.11: Functional annotation of key proteins with significant differential abundance - **chondrocytes**. Associated terms relate to gene ontology groups from which representative proteins were collected. Terms are not always significant due to small numbers of proteins but provide summary functional annotations for proteins with strong evidence for differential abundance. Central schematic (**figure legend**) indicates the different pairwise comparisons; associated with each are the KEGG canonical signalling pathways are the most significantly perturbed pathways and their predicted status: (**+**) activated, (**-**) inhibited.

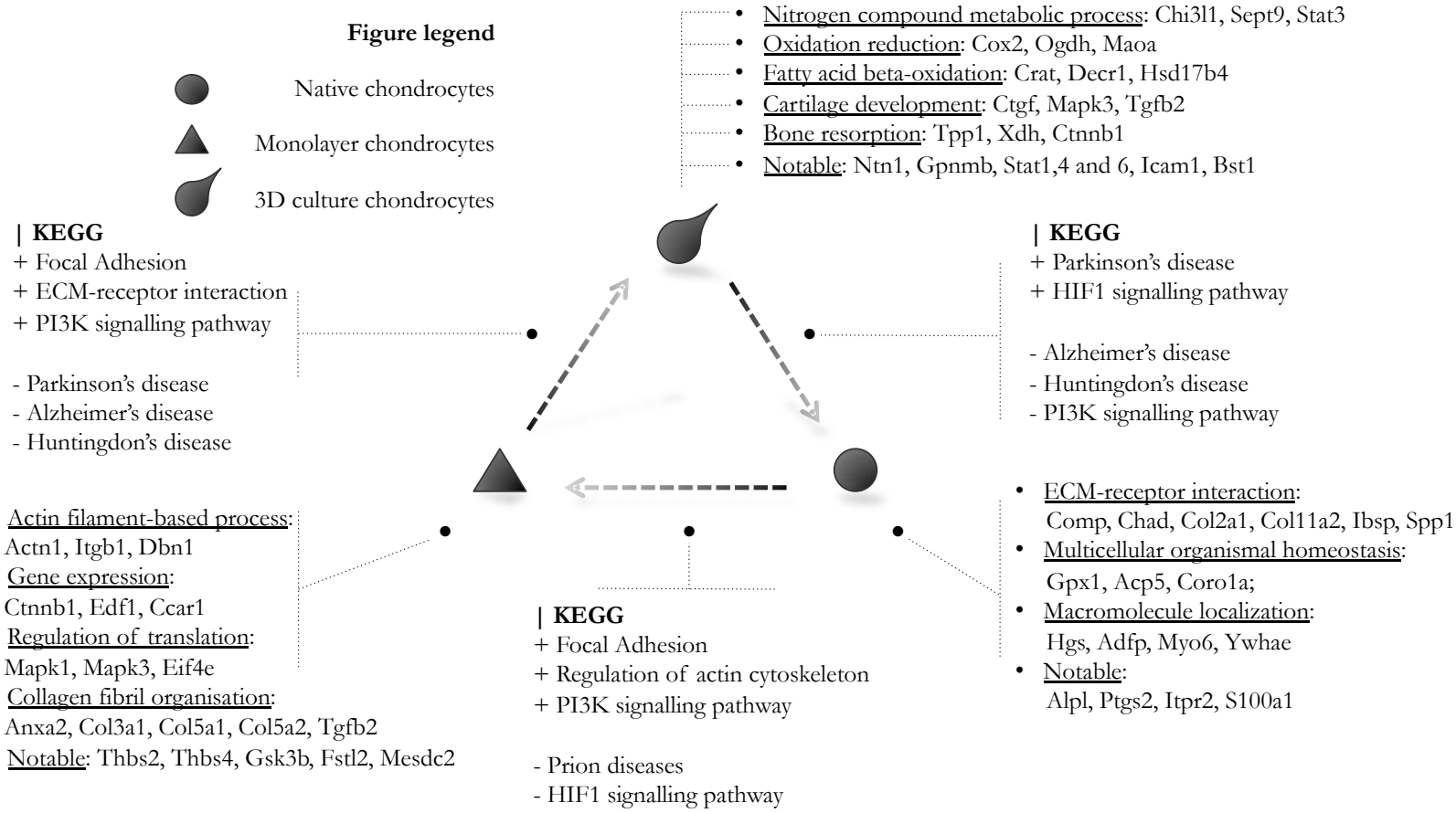
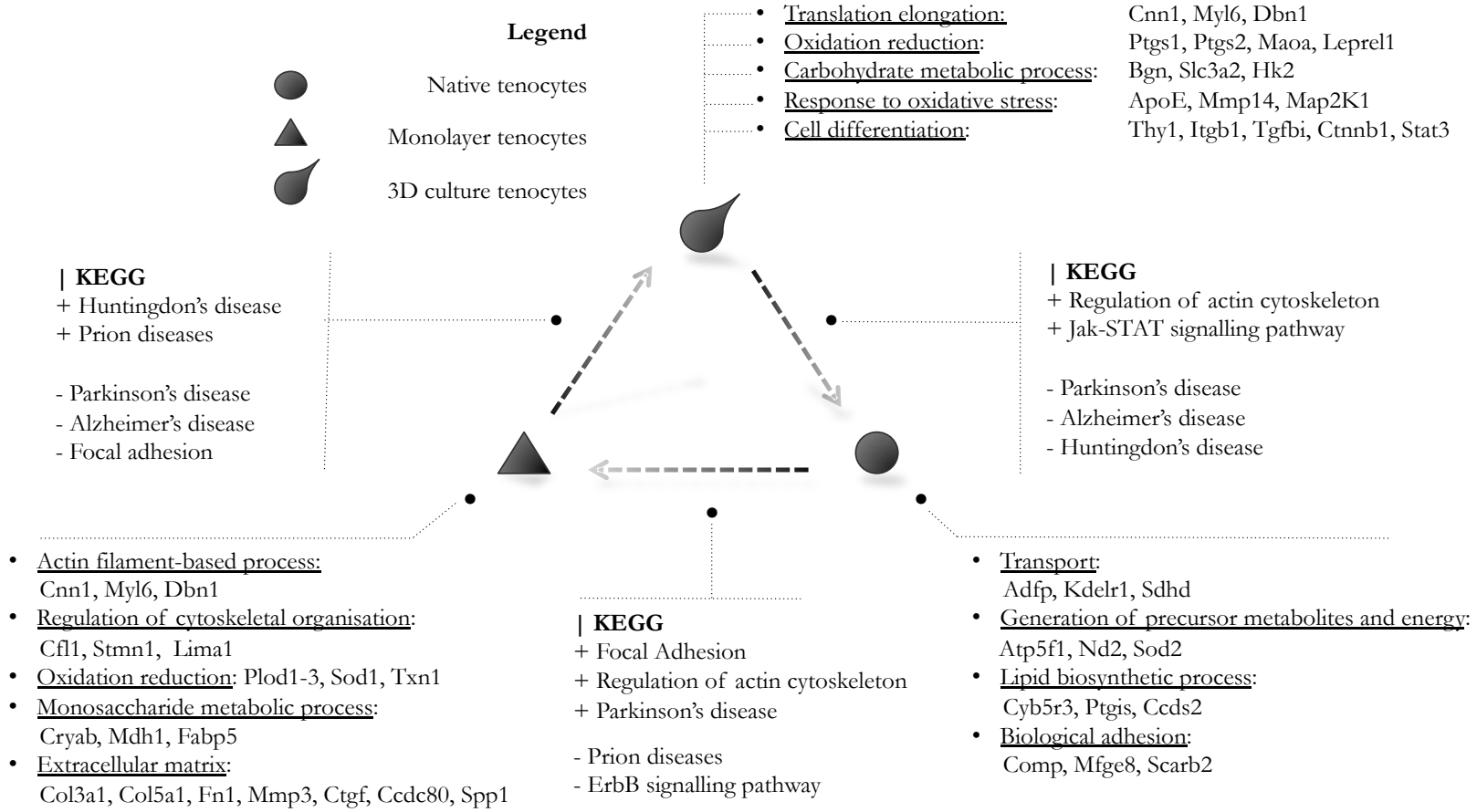


Figure 6.12: Functional annotation of key proteins with significant differential abundance - **tenocytes**. Associated terms relate to gene ontology groups from which representative proteins were collected. Terms are not always significant due to small numbers of proteins but provide summary functional annotations for proteins with strong evidence for differential abundance. Central schematic (**figure legend**) indicates the different pairwise comparisons; associated with each are the KEGG canonical signalling pathways are the most significantly perturbed pathways and their predicted status: (+) activated, (-) inhibited.



6.3.4: Moderate correlation between gene and proteomic expression data

For the native chondrocyte to dedifferentiated chondrocyte transition there were 320 elements matched across the two datasets. Overall these elements were moderately correlated, r = 0.56 ($t_{df = 318}$, p < 2.2e-16, 95% CI: 0.48-0.63), **Figure 6.13**; when only those elements with matched fold-change direction were considered this was improved, r = 0.88 ($t_{df = 240}$, p < 2.2e-16, 95% CI: 0.79-0.87). There were 42 elements more highly expressed in native chondrocytes that shared directional change in both gene expression and protein abundance surveys. These included known chondrogenic markers including collagen type 2 (COL2A1) and collagen type IX, chondroadherin (CHAD), integrin-binding sialoprotein (IBSP), matrix metalloproteinase MMP3, HAPLN1 (link protein), PRG4 (lubricin), in addition to more novel cartilage-associated proteins: calcium channel ITPR2 (inositol 1,4,5-trisphosphate receptor, type 2), and products associated with matrix mineralization, SPP1/osteopontin. Others had no known association with cartilage: dynamin (DNM1), serine protease inhibitor SERPINB1A, and glutathione-S-transferase.

For elements with higher expression in monolayer chondrocytes there were 200 that shared directional change. The following were noted: carbonic anhydrase CAR9, COL5A2, transgelin (TAGLN), the lectin LGALS1, thrombospondin 2 (THBS2), stathmin (STMN1). The only element with high anti-correlated expression was CHORDC1/morgana protein (Cysteine And Histidine-Rich Domain (CHORD) Containing 1).

For the equivalent transition for tenocytes there were 231 elements with were found in both datasets with moderate correlation in effect size, r = 0.41, ($t_{df = 329}$, p < 1.2e-10, 95% CI: 0.29-0.51), **Figure 6.14**. There were 13 elements found to be more highly expressed in native tenocytes relative to monolayer including, superoxide dismutase 2 (SOD2), serine (or cysteine) peptidase inhibitor, clade E, member 2 (SERPINE2), perilipin 2 (PLIN2) and prostaglandin I2 (prostacyclin) synthase (PTGIS). In monolayer, there were 175 elements found to be common to both analyses and shared directional change. Those with the highest correlation included: integrin, alpha 11 (ITGA11), neural cell adhesion molecule 1 (NCAM1), coiled-coil domain containing 80 (CCDC80) and Thy-1 cell surface antigen (THY1/CD90). For tenocytes CHORDC1 was also highly anti-correlated.

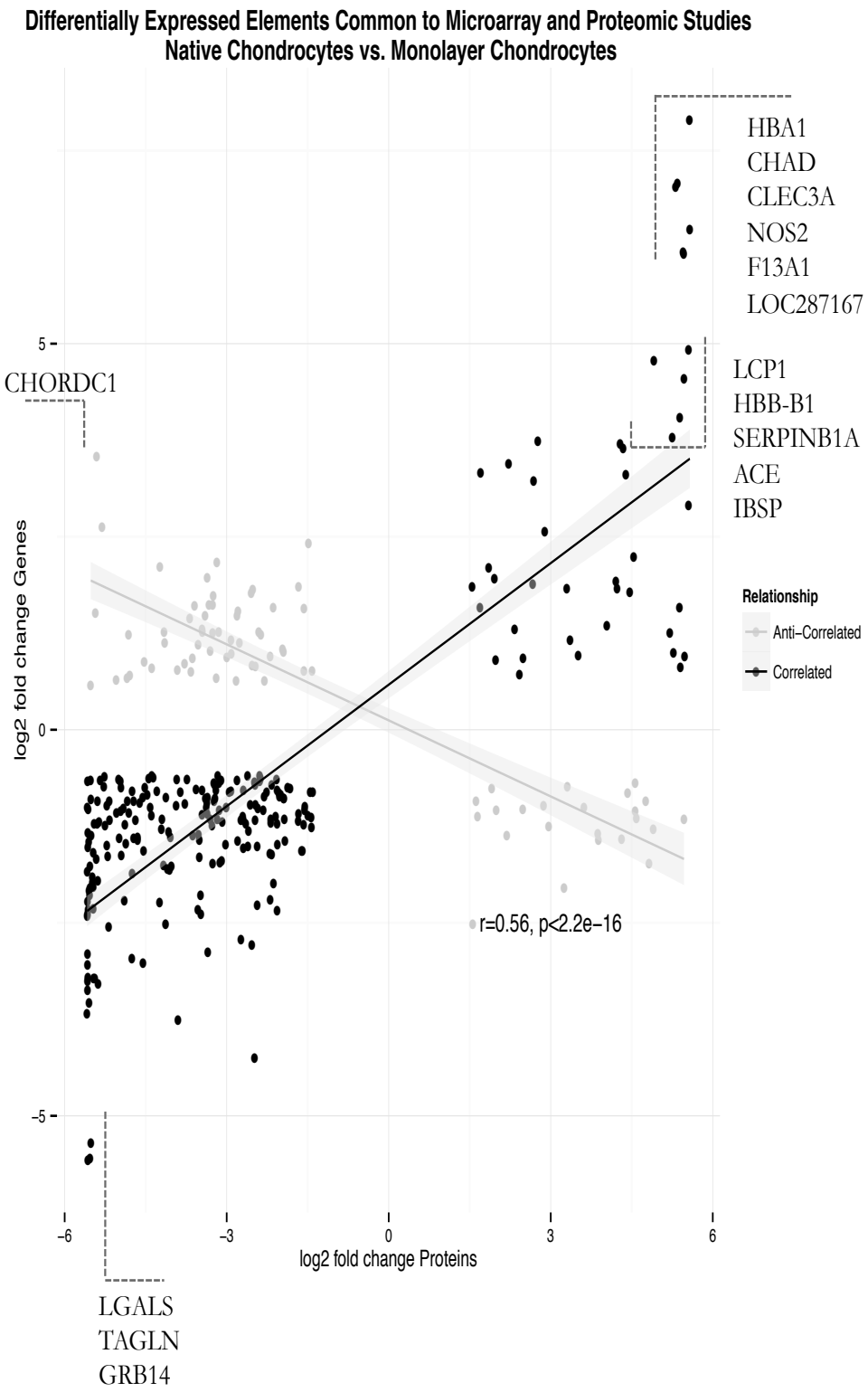


Figure 6.13: Correlation scatterplot – **chondrocytes**. Relative abundance (log_2 fold change) of proteins (x-axis) in native to monolayer transition for chondrocytes is plotted against differential expression (log_2 fold-change) for genes (y-axis) from the Affymetrix data. Dark data points indicate correlated elements (shared direction of fold change); light points indicate anti-correlated (no match) elements. Only the most highly correlated are labeled. CHORDC1 is found as an anti-correlated protein in both chondrocytes and tenocytes. Annotated lists are found in **SD6**.

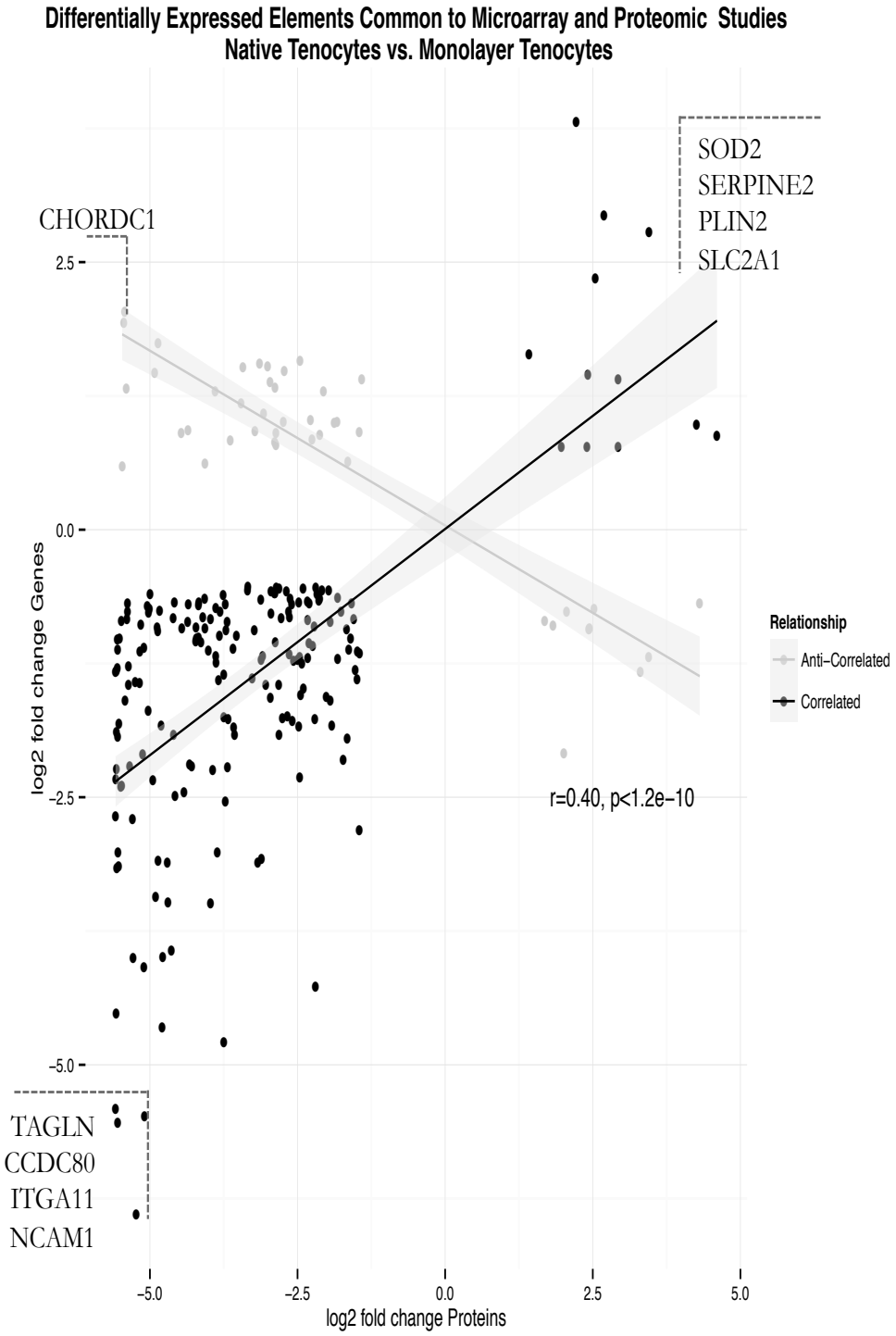


Figure 6.14: Correlation scatterplot –**tenocytes**. Relative abundance (log_2 fold change) of proteins (x-axis) in native to monolayer transition for tenocytes is plotted against differential expression (log_2 fold-change) for genes (y-axis) from the Affymetrix data. Dark data points indicate correlated elements (shared direction of fold change); light points indicate anti-correlated (no match) elements. Only the most highly correlated are labeled. Few gene and proteins from native tenocytes are found in both data sets; the majority of elements arise from monolayer data. Annotated lists are found in **SD6**.

6.3.5: Oxidative phosphorylation associated proteins are robustly expressed in three-dimensional cultures

Transcriptome profiling had found elevated levels of pro-inflammatory chemokines in three-dimensional cultures. Considering the protein abundance of these cultures found proteins strongly associated with the KEGG canonical signalling pathway 'oxidative phosphorylation' (**Figure 6.8**) in the alginate to native chondrocyte comparison, **Figure 6.15**. Proteins associated with mitochondrial dynamics, optic atrophy 1 (OPA1, C-M-T disease 1A) and cytochrome C subunits were more abundant in alginate cultures than either native or monolayer chondrocytes. Fibrin cultures shared comparable annotation for oxidative phosphorylation.

6.3.6: Common upstream regulators predicted to relate to PPAR and PI-3K signaling in alginate cultures

Analysis of alginate and native chondrocyte protein profiles using Ingenuity® Pathway Analysis defined the most significant canonical pathways as 'mitochondrial dysfunction', 'oxidative phosphorylation' and 'EIF2 signalling'. Upstream regulators were predicted for the protein profile and these including small molecule agonist of PPARα signalling, pirinixic acid (z-score = 2.19, p = 4.1e-14, 258 proteins were downstream targets of 17 regulators) **Table 6.3**, and transcription factors associated with PPARα and PI-3K signalling, **Figure 6.16**. The signal transducer and activator of transcription 3 protein (STAT3) was consistently associated with predicted mechanistic networks. In this analysis STAT3 is more highly abundant in alginate beads than in native cells; this was consistent with Illumina and Affymetrix gene expression analysis.

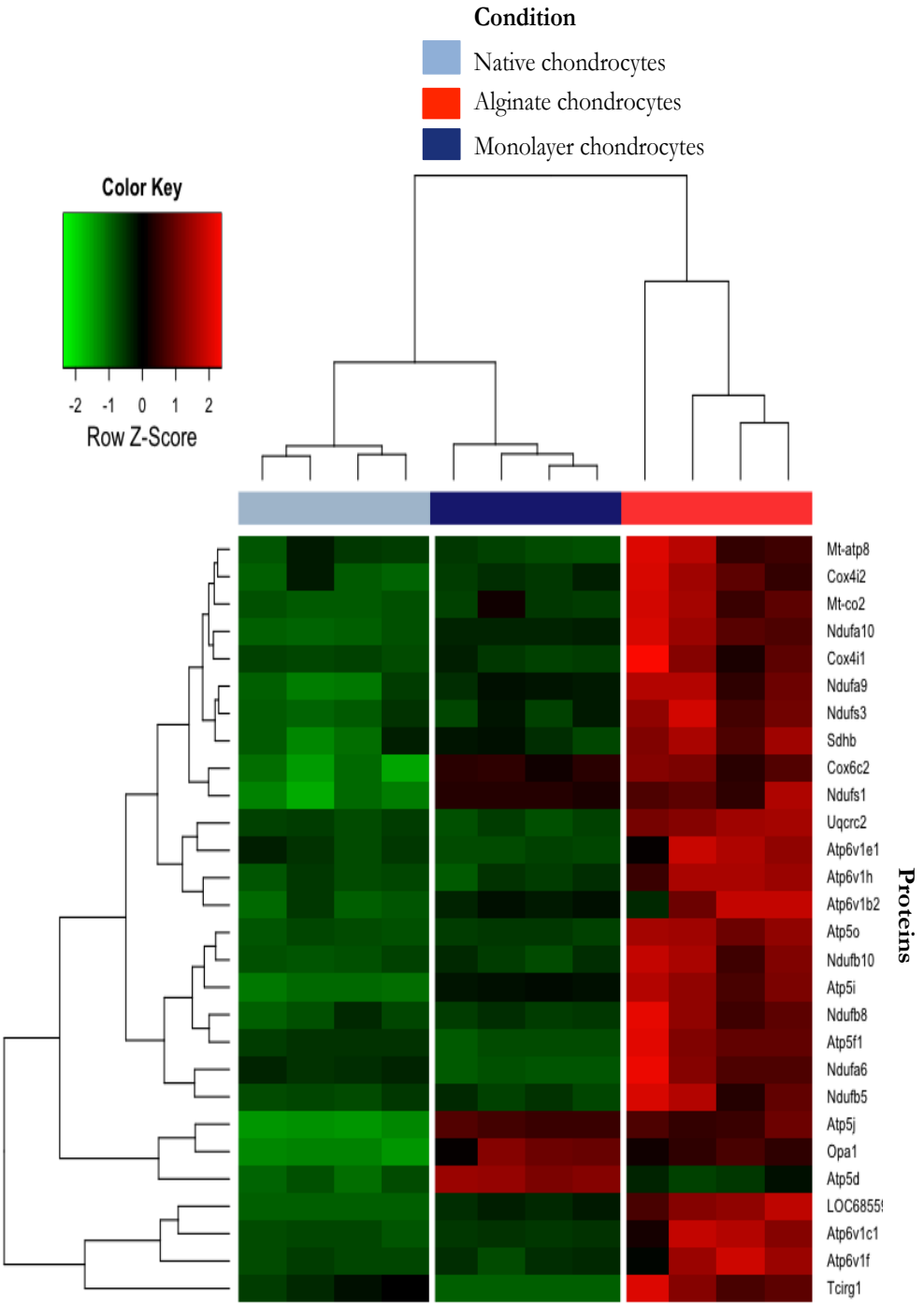


Figure 6.15 – Heatmap of $\log_2$ ratios for chondrocyte conditions using proteins enriched for the KEGG canonical signalling pathway 'oxidative phosphorylation'. Consistent with an understanding of glycolytic pathways being the main ATP source in chondrocytes from cartilage oxidative phosphorylation associated proteins are highly abundant in chondrocytes from alginate cultures.

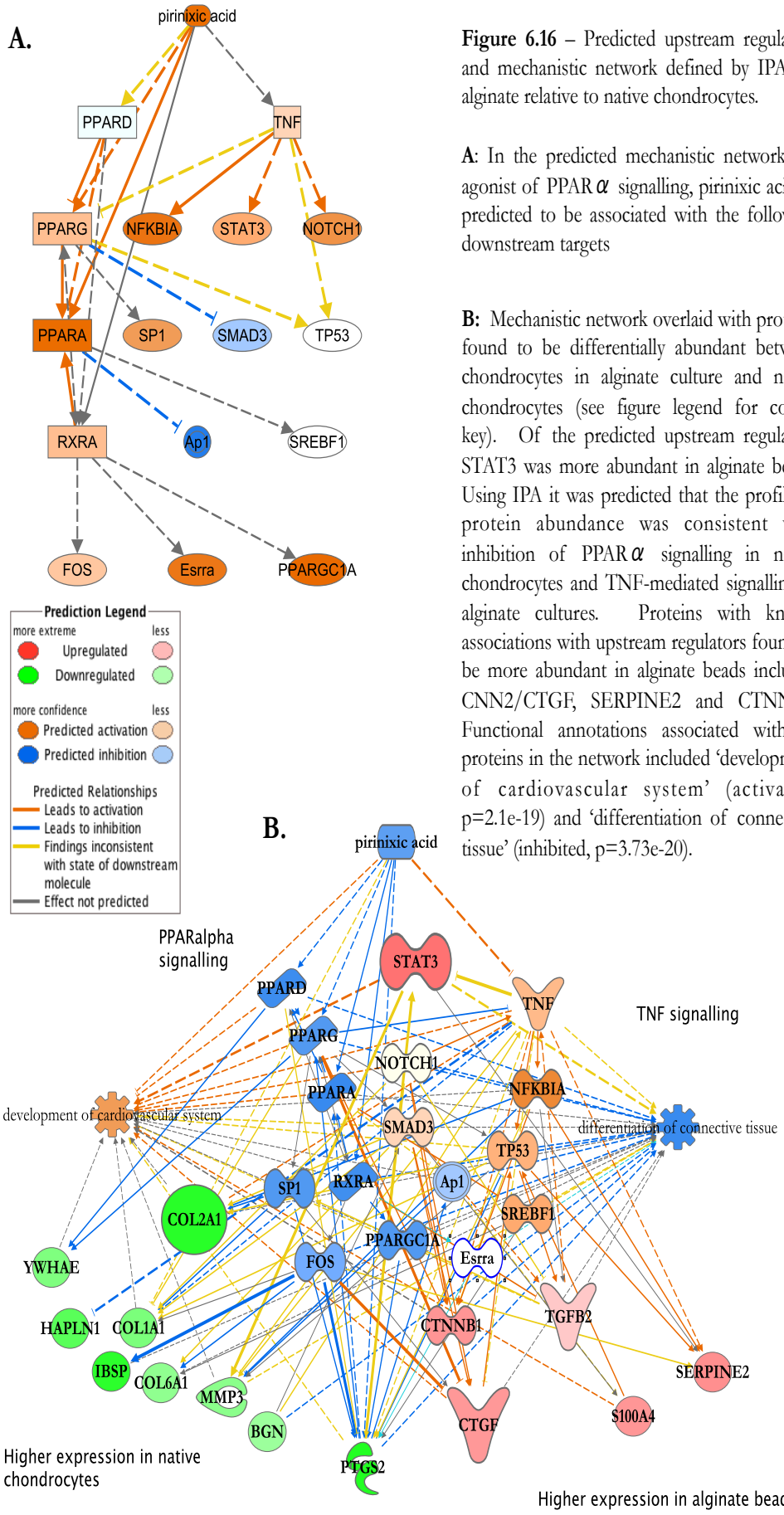


Figure 6.16 – Predicted upstream regulators and mechanistic network defined by IPA for alginate relative to native chondrocytes.

A: In the predicted mechanistic network the agonist of PPARα signalling, pirinixic acid, is predicted to be associated with the following downstream targets

B: Mechanistic network overlaid with proteins found to be differentially abundant between chondrocytes in alginate culture and native chondrocytes (see figure legend for colour key). Of the predicted upstream regulators STAT3 was more abundant in alginate beads. Using IPA it was predicted that the profile of protein abundance was consistent with inhibition of PPARα signalling in native chondrocytes and TNF-mediated signalling in alginate cultures. Proteins with known associations with upstream regulators found to be more abundant in alginate beads included CNN2/CTGF, SERPINE2 and CTNNB1. Functional annotations associated with all proteins in the network included 'development of cardiovascular system' (activated, p=2.1e-19) and 'differentiation of connective tissue' (inhibited, p=3.73e-20).

		Comparison	Pirinixic acid	D3T
A.	Chondrocytes	Native to monolayer	-4.43 (p=1.23e-11)	-6.04 (p=2.23e-37)
		Monolayer to alginate	3.37 (p=1.23e-23)	4.69 (p=3.4e-27)
		Alginate to native	2.19 (p=4.8e-14)	Not represented
B.	Tenocytes	Native to monolayer	-5.05 (p=2.02e-12)	-7.3 (p=6.52e-40)
		Monolayer to fibrin	1.4 (p=2.81e-16)	1.9 (p=5.9e-20)
		Fibrin to native	3.06 (p=4.5e-14)	6.01 (p=2.24e-29)

Table 6.3: Chemical reagents (columns) used by Ingenuity Pathway Analysis as proxy regulators of signalling pathways based upon differentially abundant proteins in **A**) chondrocytes and **B**) tenocyte comparisons (rows). Cells represent activation z-score and overlap p-value. The actions of the PPAR-alpha agonist pirinixic acid and the anti-oxidant 1,2 dithiol-3-thione (D3T) are both shown to be consistent with the protein abundance profiles in monolayer and three-dimensional cultures (excluding alginate to native chondrocyte comparison).

6.4: Discussion

6.4.1: Study design and limitations

In a review of cartilage proteomic surveys Hsueh, *et al* (2014) found that between 14 and 814 proteins could be identified, depending on the methods used and whether whole tissue or *in vitro* chondrocytes were profiled (Hsueh, Önnerfjord et al. 2014). Relative to these studies the profiling of matrix-depleted native cells in this study resulted in higher numbers of differentially abundant proteins, greater than 1800 across both chondrocytes and tenocytes. This exemplifies the trade-off between the depth of coverage achieved with mass spectrometry and the accurate representation of the native state. Additional levels of statistical analysis and *post-hoc* filtering were employed to ensure that only the most robust changes were retained for differential abundance analysis. The large number of confidently identified and differentially abundant proteins permitted the use of over-representation analysis for functional annotation and pathway topology analysis methods and facilitated significant descriptors to be derived from them.

This study was unusual in that it surveyed both gene expression and protein abundance in parallel with sample collection and culture periods temporally matched. Consistent sample handling, awareness of cell culture media contaminants (phenol red), application of multiple wash phases, use of glass and low-affinity plastic storage tubes, the depletion of abundant, anionic matrix and local technical expertise were all practices that may have contributed to the identification of large numbers of proteins and the high intra-sample correlation in proteomic profiles.

The principal limitation was the number of cells available for analysis, especially for native tenocytes and three-dimensional culture samples. Although methodologies exist for the co-extraction of RNA and protein the low cell yield from samples in the preliminary study indicated that paired biological replicates would not be possible. Furthermore, for the same wet weight, tendon consistently yielded fewer cells than cartilage and required special attention to ensure optimal digestion of fibrillar material (Ritty, Roth et al. 2003). Sample handling in this study was not optimal with cells frozen as suspensions in PBS with protease inhibitors rather than as cell pellets. This did influence the concentration of protein that was loaded on the LC for tenocytes, but did not ultimately appear to have adversely influenced the number of proteins identified with comparable numbers identified in both native cell groups.

The depletion of extra- and peri-cellular matrix (PCM) results in a loss of information relating to proteins with key roles in cartilage and tendon homeostasis, for example integrins and other plasma membrane proteins (Iliopoulos, Gkretsi et al. 2010), though some were represented in this analysis. An assumption of PCM depletion was based upon a particle-exclusion test, which although cheap and quick to perform limited assessment to individual cells tested from aliquots in preliminary studies, i.e. it gave no indication as to the overall sample reduction in PCM components for the final analysis. Validation of type VI collagen depletion by Western blotting or loss of Alcian blue staining for polysulphated proteoglycans would have provided more robust qualitative indicators of PCM depletion. The low fold-changes associated with extra-cellular matrix proteins and the lack of ECM-associated annotations from differentially

abundant proteins would suggest that depletion approach was effective, however all three collagen type VI alpha chains were found to be differentially abundant.

This study utilized the linear Bayes moderated t-test within `Limma` (Smyth 2005) applied to the gene expression data to provide a comparable way of filtering the data and providing a confidence value to the identification of differentially abundant proteins. This approach is now relatively commonly used in the proteomics literature (Katayama, Paczesny et al. 2009, Ting, Cowley et al. 2009, Schwämmle, León et al. 2013, Wanner, Subbaiah et al. 2013) and, as such, is a valid methodology within this context. Absolute or labeled-relative quantification would be an obvious step to validate differential abundance in proteins of interest.

6.4.2: Oxidative phosphorylation in three dimensional culture systems

Given the considerable overlap in protein surveys from different cells sources and conditions there is confidence in the significantly enriched functional annotations presented. 'Oxidative phosphorylation' and 'oxidation-reduction' were the consistent descriptors for three-dimensional model cultures as compared to 'actin filament-based processes' for monolayer cultures. In tandem with these findings across all comparisons canonical signalling pathways associated with neurodegenerative conditions were strongly enriched and shown to have differential, although conflicted, activation status. Furthermore, proteins associated with metabolic processes – carbohydrate and fatty acid metabolism - abounded.

How these findings are interpreted is conditional upon the view taken of the nature of cells in three-dimensional cultures, i.e. whether they represent a regenerative, re-differentiating population, or the representation of a degenerative state. These findings also have relevance to the use of mesenchymal stem cells for

cartilage directed-differentiation studies or defining populations of adult tissue stem-cells. On this basis the relevance of oxidative phosphorylation (OXPHOS) to musculoskeletal biology should be considered.

OXPHOS is the metabolic process of ATP synthesis in mitochondria at the expense of respired oxygen. Redox reactions, where electrons are transferred from donors (NADH from the citric acid cycle) to acceptors, for example oxygen (terminal acceptor), along the electron transport chain release energy used to reform ATP from ADP. The addition of a phosphate group to ADP is driven by proton gradients created by the harnessing of energy released in the electron transport chain to drive proton pumps (Alberts, Bray et al. 1998). The terminal electron acceptor in the respiratory chain is oxygen, which with the addition of four electron and two protons forms water. Although this is efficient a small number of electrons will partially reduce oxygen, with the addition of only one or two electrons, to produce harmful reactive oxygen species superoxide and peroxide.

In both the gene and protein surveys of native, monolayer and model culture conditions over-representation and pathway topology analysis have implicated the KEGG canonical pathways for the neurodegenerative diseases: Parkinson's, Huntington's and Alzheimer's. Initially this does not seem intuitive to the understanding of musculoskeletal cell biology, however, given the considerable evidence supporting mitochondrial dysfunction in these three conditions (Correia, Santos et al. 2012, Johri and Beal 2012) the annotation of these proteomics data set with neurodegenerative pathways is a consequence of the component proteins having associations with mitochondrial homeostasis and/or dysfunction. This would indicate the importance of OXPHOS and redox balance for homeostatic

mechanisms in chondrocytes and tenocytes. For example, this study presents evidence for high abundance of proteins associated with mitochondrial dynamics and fission/fusion, OPA1 (optic atrophy 1) in model cultures and MFN2, mitofusin-2, (associated with Charcot-Marie-Tooth type 2A) in fibrin cultures.

OXPHOS in chondrogenesis and regeneration

The process of OXPHOS has relevance to the condensation and differentiation of chondrocytes. Using a bioluminescent monitoring system Kwon, *et al* (2012) found that ATP oscillations were dependent on glycolysis, oxidative phosphorylation and Ca^{2+} levels, and were synchronized by gap junctions across chondrocytes during condensation (Kwon, Ohmiya et al. 2012). This condensation was retarded by blockade of the ATP oscillations. The authors concluded that this type of synchronized oscillatory activity during condensation would be relevant to the periodic secretion of adhesion molecules and ECM during condensation.

Chondrocytes are known to reside in hypoxic conditions (Pfander and Gelse 2007), have few mitochondria, and generate the majority of their ATP (> 90%) by anaerobic glycolysis rather than OXPHOS (Otte 1991, Martin, Martini et al. 2012) yet hypoxia is also known to promote maintenance of stem cells, including MSCs, and these are more reliant on glycolysis than OXPHOS. This may be a mechanism for the prevention of oxidative stress-induced senescence occurring by OXPHOS under normoxia (Ito and Suda 2014). It is not clear from the data arising in this study whether OXPHOS activity is a response to culture conditions alone and/or represents changes in the differentiation status of cells in three-dimensional cultures. Under normoxic conditions mesenchymal stem cells have been shown to significantly reduce oxygen consumption when differentiating

towards a chondrocytic phenotype in pellet cultures, in contrast to osteoblastic differentiation (Pattappa, Heywood et al. 2011). The dedifferentiated chondrocyte, in contrast, shows an increased reliance on OXPHOS with increasing periods of time in culture (Heywood and Lee 2008), with expansion under hypoxic and low-glucose conditions enhancing the subsequent differentiation capacity of monolayer-expanded cells in pellet culture (Heywood, Nalesso et al. 2014).

Additionally, PI-3K/Akt/mTOR has a 'nutrient sensing' activity in response to glucose and amino acids. In the maintenance of stem cell pools equilibrium in metabolic pathways is required where mitochondria in stem cells are relatively quiescent and the reliance is on glycolytic pathways. A change to oxidative phosphorylation is associated with impaired stem cell function and differentiation (Ito and Suda 2014). Where the redox rheostat is finely balanced to maintain self-renewal and inhibit differentiation comparable mechanisms may be relevant in dedifferentiation/re-differentiation mechanisms for musculoskeletal cells.

PI-3K/Akt signalling and OXPHOS

In **Chapter 3** pathway topology analysis of the gene expression profiles of the same conditions implicated PI-3K/Akt signalling pathway as the over-arching regulatory mechanism. In the proteomic profiles this remained a significantly perturbed pathway for chondrocytes, but the predicted activation statuses were contradictory. The role of PI-3K/Akt signalling is not at odds with findings of oxidative phosphorylation pathways in the proteomic survey. The production of ROS is enhanced by PI-3K/Akt signalling through the repression of FOXO-mediated responses (Ito and Suda 2014). As discussed in **Chapter 3** oxidative stress appears to inhibit IGF1 induction of matrix synthesis by failing to activate

Akt by phosphorylation (Yin, Park et al. 2009) via PTEN (Iwasa, Hayashi et al. 2014).

OXPHOS, oxidative stress and OA

In **Chapter 3 and 5** the high expression of chemokines in model cultures was presented. In this proteomic survey few are represented in the differential abundance analysis – this is perhaps unsurprising given that IL-6, for example, is largely secreted and found at high levels in culture media (Tsuchida, Beekhuizen et al. 2012). The question arises as to whether oxidative stress induced by OXPHOS promotes a pro-inflammatory cytokine signature or whether cytokine release is within the re-differentiation mechanism of cells in model culture. Recently Cao and colleagues (2013) found that mitochondria in cultured murine chondrocytes, and *in situ* within transgenic mice, responded to IL-1β or TNFα with 'super-oxide flashes' 2-5 fold more frequently than in quiescent chondrocytes (Cao, Zhang et al. 2013)

There is evidence associating mitochondrial dysfunction and oxidative stress with chondropathies and age-related degeneration in general. The density of mitochondria decreases with increasing depth from the surface of articular cartilage in line with an understanding of decreasing oxygen tension from ~5% to < 1% from the superficial to deep layers (Blanco, Rego et al. 2011, Cao, Zhang et al. 2013). Osteoarthritis is associated with the elevated production of ROS by chondrocytes and inhibition of respiratory chain complexes III or V reduces ROS levels, but inhibition results in the induction of pro-inflammatory cytokine production (Blanco, Rego et al. 2011).

Expression of superoxide dismutase 2 (mitochondrial), SOD2, is significantly decreased in OA chondrocytes and cartilage (Blanco, Rego et al. 2011). Using an MS time-of-flight approach to 2D DIGE resolved proteins from human articular chondrocytes (HACS) Ruiz-Romero, *et al* (2009) defined a significant decrease in the levels of SOD2 in osteoarthritis HACS, with a concomitant higher expression of ROS and TRAP1, a heat-shock protein 90 family-member with ROS antagonistic activity (Ruiz-Romero, Calamia et al. 2009); further work found TRAP1 protein to be more highly expressed under hypoxic conditions (Ruiz-Romero, Calamia et al. 2010) and may act as a molecular switch between OXPHOS and glycolysis (Yoshida, Tsutsumi et al. 2013). In this study SOD2 is highly abundant in alginate and fibrin cultures relative to native cells and TRAP1 is expressed at higher levels in alginate cultures in normoxic conditions relative to native cells, but not in fibrin cultures. This would imply an anti-oxidant response is present in alginate-encapsulated chondrocytes and does not necessarily indicate that there is a redox imbalance.

Summary

On the basis of protein abundance, results presented in this study indicate that oxidative phosphorylation contributes to ATP synthesis in model cultures. The considerable qualitative overlap in confidently identified proteins in both fibrin and alginate cultures suggest that this is common to both systems.

Given the evidence for mitochondrial dysfunction in cartilage pathology and a potential role in cellular differentiation and self-renewal of stem cells the metabolic profile of cells in organo-typic cultures should be validated and standardised. Recent studies indicate dynamic activity in chondrocytes in terms of ATP production and super-oxide concentrations. This temporal dynamic should also

be considered in model culture systems. If alginate cultures represent a superoxide and pro-inflammatory soup they may represent a better model of osteoarthritis than of the physiological state.

6.4.3: Proteomic and gene expression correlations

As outlined in the introduction the correlation between mRNA and protein levels is poorly correlated representing as it does early and late stages of an observed regulatory event with no appreciation of the complex modulation occurring concurrently (Payne 2015). Tian, *et al* (2004) reported that differential expression analysis of mRNA would account for no more than 40% of the variation in protein expression; this uncoupling of the two expression profiles highlighting the importance of post-translational regulatory mechanisms (Tian, Stepaniants et al. 2004).

Moderate correlation is shown between differentially expressed elements from proteomics and transcriptomic studies from the same experimental groups using the same study design and methods. It is worth considering this objectively – firstly, the veracity of these findings could be accepted and the conclusion made that the identification and correlation of the same proteins in chondrocytes and tenocytes indicated some equivalence in processes. The alternative view is that there is bias in the lists of identified proteins, as a consequence of the nature of tryptic digestion and peptide behavior in the MS, and so, also bias in the differential abundance findings. Further validation of correlations between gene and protein levels of the serine protease inhibitors SERPINE1 and SERPINB1A, IBSP and SOD2 in native chondrocytes or tenocytes by qPCR and Western blotting would be required. Consideration of parallel miRNA studies, metabolomics, protein phosphorylation status and the definition of dynamic

cellular localization in association with gene and protein expression profiling would assist interpretation of future experiments i.e. whole cell modeling (Karr, Sanghvi et al. 2012), however, integration of multiple data sources and not defining correlation 'correctness' should be the focus of future research (Payne 2015).

Utility of anti-correlated mRNA and protein expression

Even given the appreciation of the discordant relationship between gene and protein expression there is still utility in this type of correlation analysis. Anti-correlated elements (where the fold-change direction is not shared in mRNA and protein expression profiling) are of interest as they may disclose additional information representing post-translational regulatory mechanisms (Tian, Stepaniants et al. 2004), e.g. miRNA. Payne (2014) notes that these differences are fundamentally related to regulatory mechanisms rather than measurement errors (Payne 2015). Here the highly anti-correlated expression of morgana protein/CHORDC1/Chp-1, a ROCK inhibitor (Ferretti, Palumbo et al. 2010) and HSP90 interacting protein (Wu, Luo et al. 2005), is reported between mRNA and protein expression in chondrocytes and tenocytes in the native to monolayer comparison. In both scenarios high mRNA expression in native cells is associated with low protein abundance. There are no publications investigating the role of CHORDC1 in cartilage or tendon physiology and so it should be considered in a wider context. Recently the proto-oncogene activity of morgana protein has been associated with PTEN destabilization, through the inhibition of ROCK (Rho family of GTPases), promoting PI-3K/Akt signalling (Fusella, Ferretti et al. 2014). Modulation of ROCK is well described as being relevant to the differentiation status of chondrocytes and to cartilage function. Inhibition of ROCK has effects

on SOX9 activation and the suppresses dedifferentiation in chondrocytes (Woods and Beier 2006, Matsumoto, Furumatsu et al. 2012). Mechnotransduction through RhoA/ROCK is also a key modulator of CNN2/CTGF and TGFβ-Smad signaling with inhibitors of ROCK inhibiting CNN2/CTGF expression (Chaqour and Goppelt-Struebe 2006). Therefore, although the correlation of expression changes between mRNA and protein may provide some level of validation of findings the anti-correlated elements are also sources of further complexity in our wider understanding of cartilage physiology. In this case there is rationale for the further investigation of CHORDC1 mRNA and protein expression in relation to a mechanistic network involving PI-3K/Akt, RhoA/ROCK, CCN-family and TGFβ-Smad signaling in chondrocyte differentiation status.

6.4.4: PPAR signalling

A predicted role for peroxisome proliferator-activated receptor (PPAR) signalling in native chondrocytes is introduced for the first time in this thesis. Using the PPARα agonist pirinxic acid as a proxy for activation of this pathway Ingenuity® Pathway Analysis predicted inhibition where the protein profile was associated with high abundance in native cells compared to monolayer; there was predicted activation in both monolayer and three-dimensional cultures for both chondrocytes and tenocytes. With Ingenuity® analysis the activation of PPAR signalling follows the prediction of D3T-mediated effects. D3T is an anti-oxidant, which stabilizes the NRF2 transcription factor involved in responses to oxidative stress (Kwak, Itoh et al. 2001).

Peroxisomes are membrane-bound vesicles for the storage and degradation of reactive chemicals and hydrogen peroxide. They contain oxidative enzymes and a major function is the breakdown of long chain fatty acids by beta-oxidation

(Alberts, Bray et al. 1998). Peroxisome proliferator-activated receptors were identified as ligand-inducible transcription factors that induced peroxisome proliferation (Dreyer, Krey et al. 1992). PPAR isotypes (α, β/δ, or γ isotypes) have different gene targets and distinct tissue distributions (Gorniak 2014) with roles in lipid metabolism, inflammation, insulin sensitivity, metabolism (Ahmadian, Suh et al. 2013) and mitochondrial biogenesis, acting in heterodimer partnership with retinoid X receptor-α (RXRα)

With respect to cartilage research the focus has been on the PPARγ transcription factor; most recently PPARγ knock-out mice have been shown to develop spontaneous osteoarthritis (Vasheghani, Monemdjou et al. 2013) and impaired endochondral ossification (Monemdjou, Vasheghani et al. 2012) suggesting that PPARγ is a critical regulator of cartilage health and development. A PPARγ agonist, given orally to mice, was demonstrated to diminish the severity of collagen-induced arthritis (Tomita, Kakiuchi et al. 2006). In an *in vivo* study of rat chondrocytes encapsulated in alginate beads, however, the application of various PPAR agonists was shown to inhibit the TGF-β1 stimulating effects on prostaglandin synthesis suggesting deleterious effects of some PPAR agonists (Poleni, Bianchi et al. 2007).

Although separately oxidative stress, mitochondrial dysfunction and PPAR signalling have all been investigated in cartilage, as described above, an integrated understanding of these mechanisms in cartilage physiology is not apparent in the literature. Given the supporting evidence there is a rationale for further research relating to the contribution of PPAR-signalling in cartilage homeostasis and disease.

6.4.5: Summary

A proteomic discovery profiling study of chondrocytes and tenocytes, employing a label-free relative quantification approach, demonstrated moderate associations with a gene expression study performed in parallel. This was in line with current understanding of the expected correlation between mRNA and protein in expression studies. Functional annotation, pathway topology analysis and prediction of upstream regulators was comparable across chondrocyte and tenocyte studies implicating oxidative phosphorylation, oxidative stress and PPAR signalling as mechanisms involved within *in vitro* culture systems. These findings were broadly in line with those found within the Affymetrix gene expression studies. Further validation focused on validating the metabolic activity of cells in three-dimensional cultures is required.

References

Ahmadian, M., J. M. Suh, N. Hah, C. Liddle, A. R. Atkins, M. Downes and R. M. Evans (2013). "PPAR[gamma] signaling and metabolism: the good, the bad and the future." Nature Medicine **99**(5): 557-566.

Alberts, B., D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter (1998). Energy generation in mitochondria and chloroplasts. Essential Cell Biology: An introduction to the molecular biology of the cell, Garland Publishing, Inc.**:** 407-446.

Bantscheff, M., S. Lemeer, M. Savitski and B. Kuster (2012). "Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present." Analytical and Bioanalytical Chemistry **404**(4): 939-965.

Bantscheff, M., M. Schirle, G. Sweetman, J. Rick and B. Kuster (2007). "Quantitative mass spectrometry in proteomics: a critical review." Analytical and Bioanalytical Chemistry **389**(4): 1017-1031.

Beck, M., M. Claassen and R. Aebersold (2011). "Comprehensive proteomics." Current Opinion in Biotechnology **22**(1): 3-8.

Benninghoff, A. (1925). "Form und Bau der Gelenkknorpel in ihren Beziehungen zur Funktion. Zweiter tell: der aufbau des gelenkknorpels in seinen beziehungen zur funktion." Zeitschrift für Zellforschung und Mikroskopische Anatomie **2**(5): 783-862.

Blanco, F., I. Rego and C. Ruiz-Romero (2011). "The role of mitochondria in osteoarthritis." Nature Reviews: Rheumatology **7**(3): 161-169.

Boschetti, E. and P. G. Righetti (2008). "The ProteoMiner in the proteomic arena: a non-depleting tool for discovering low-abundance species." Journal of Proteomics **71**(3): 255-264.

Cao, Y., X. Zhang, W. Shang, J. Xu, X. Wang, X. Hu, Y. Ao and H. Cheng (2013). "Proinflammatory Cytokines Stimulate Mitochondrial Superoxide Flashes in Articular Chondrocytes In Vitro and In Situ." PLoS ONE **8**(6): e66444.

Chaqour, B. and M. Goppelt-Struebe (2006). "Mechanical regulation of the Cyr61/CCN1 and CTGF/CCN2 proteins." The FEBS Journal **273**(16): 3639-3649.

Correia, S., R. Santos, G. Perry, X. Zhu, P. Moreira and M. Smith (2012). "Mitochondrial importance in Alzheimer's, Huntington's and Parkinson's diseases." Advances in Experimental Medicine and Biology **724**: 205-221.

de Sousa Abreu, R., L. Penalva, E. Marcotte and C. Vogel (2009). "Global signatures of protein and mRNA expression levels." Molecular bioSystems **5**(12): 1512-1526.

Deutsch, E., H. Lam and R. Aebersold (2008). "Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics." Physiological Genomics **33**(1): 18-25.

Dreyer, C., G. Krey, H. Keller, F. Givel, G. Helftenbein and W. Wahli (1992). "Control of the peroxisomal beta-oxidation pathway by a novel family of nuclear hormone receptors." Cell **68**(5): 879-887.

Ferretti, R., V. Palumbo, A. Di Savino, S. Velasco, M. Sbroggiò, P. Sportoletti, L. Micale, E. Turco, L. Silengo, G. Palumbo, E. Hirsch, J. Teruya-Feldstein, S. Bonaccorsi, P. P. Pandolfi, M. Gatti, G. Tarone and M. Brancaccio (2010). "Morgana/chp-1, a ROCK inhibitor involved in centrosome duplication and tumorigenesis." Developmental Cell **18**(3): 486-495.

Fusella, F., R. Ferretti, D. Recupero, S. Rocca, A. Di Savino, G. Tornillo, L. Silengo, E. Turco, S. Cabodi, P. Provero, P. P. Pandolfi, A. Sapino, G. Tarone and M. Brancaccio (2014). "Morgana acts as a proto-oncogene through inhibition of a ROCK-PTEN pathway." The Journal of Pathology **234**(2): 152-163.

Goebel-Stengel, M., A. Stengel, Y. Taché and J. Reeve (2011). "The importance of using the optimal plasticware and glassware in studies involving peptides." Analytical Biochemistry **414**(1): 38-46.

Gorniak, B. (2014). "Peroxisome proliferator-activated receptors and their ligands: nutritional and clinical implications - a review." Nutrition Journal **13**(1): 17.

Heywood, H., G. Nalesso, D. Lee and F. Dell'accio (2014). "Culture expansion in low-glucose conditions preserves chondrocyte differentiation and enhances their subsequent capacity to form cartilage tissue in three-dimensional culture." BioResearch Open Access **3**(1): 9-18.

Heywood, H. K. and D. A. Lee (2008). "Monolayer expansion induces an oxidative metabolism and ROS in chondrocytes." Biochemical and Biophysical Research Communications **373**(2): 224-229.

Hsueh, M.-F., P. Önnerfjord and V. B. Kraus (2014). "Biomarkers and proteomic analysis of osteoarthritis." Matrix Biology **39**: 56-66.

Iliopoulos, D., V. Gkretsi and A. Tsezou (2010). "Proteomics of osteoarthritic chondrocytes and cartilage." Expert Review of Proteomics **7**(5): 749-760.

Ito, K. and T. Suda (2014). "Metabolic requirements for the maintenance of self-renewing stem cells." Nature Reviews. Molecular Cell Biology **15**(4): 243-256.

Iwasa, K., S. Hayashi, T. Fujishiro, N. Kanzaki, S. Hashimoto, S. Sakata, N. Chinzei, T. Nishiyama, R. Kuroda and M. Kurosaka (2014). "PTEN regulates matrix synthesis in adult human chondrocytes under oxidative stress." Journal of Orthopaedic Research **32**(2): 231-237.

Johri, A. and F. Beal (2012). "Mitochondrial Dysfunction in Neurodegenerative Diseases." Journal of Pharmacology and Experimental Therapeutics **342**(3): 619-630.

Karr, Jonathan R., Jayodita C. Sanghvi, Derek N. Macklin, Miriam V. Gutschow, Jared M. Jacobs, B. Bolival, Jr., N. Assad-Garcia, John I. Glass and Markus W. Covert (2012). "A Whole-Cell Computational Model Predicts Phenotype from Genotype." Cell **150**(2): 389-401.

Katayama, H., S. Paczesny, R. Prentice, A. Aragaki, V. Faca, S. Pitteri, Q. Zhang, H. Wang, M. Silva, J. Kennedy, J. Rossouw, R. Jackson, J. Hsia, R. Chlebowski, J. Manson and S. Hanash (2009). "Application of serum proteomics to the Women's Health Initiative conjugated equine estrogens trial reveals a multitude of effects relevant to clinical findings." Genome Medicine **1**(4): 47.

Kielty, C. M., M. Lees, C. A. Shuttleworth and D. Woolley (1993). "Catabolism of intact type VI collagen microfibrils: susceptibility to degradation by serine proteinases." Biochemical and Biophysical Research Communications **191**(3): 1230-1236.

Kraut, A., M. Marcellin, A. Adrait, L. Kuhn, M. Louwagie, S. Kieffer-Jaquinod, D. Lebert, C. D. Masselon, A. Dupuis, C. Bruley, M. Jaquinod, J. Garin and M. Gallagher-Gambarelli (2009). "Peptide Storage: Are You Getting the Best Return on Your Investment? Defining Optimal Storage Conditions for Proteomics Samples." Journal of Proteome Research **8**(7): 3778-3785.

Kuettner, K. E., B. U. Pauli, G. Gall, V. A. Memoli and R. K. Schenk (1982). "Synthesis of cartilage matrix by mammalian chondrocytes in vitro. I. Isolation, culture characteristics, and morphology." The Journal of Cell Biology **93**(3): 743-750.

Kwak, M. K., K. Itoh, M. Yamamoto, T. R. Sutter and T. W. Kensler (2001). "Role of transcription factor Nrf2 in the induction of hepatic phase 2 and antioxidative enzymes in vivo by the cancer chemoprotective agent, 3H-1, 2-dimethiole-3-thione." Molecular Medicine (Cambridge, Mass.) **7**(2): 135-145.

Kwon, H. J., Y. Ohmiya, K. i. Honma, S. Honma, T. Nagai, K. Saito and K. Yasuda (2012). "Synchronized ATP oscillations have a critical role in prechondrogenic condensation during chondrogenesis." Cell Death & Disease **3**(3): e278.

Lee, G. M., C. A. Poole, S. S. Kelley, J. Chang and B. Caterson (1997). "Isolated chondrons: a viable alternative for studies of chondrocyte metabolism in vitro." Osteoarthritis and Cartilage **5**(4): 261-274.

Lin, H., L. He and B. Ma (2013). "A combinatorial approach to the peptide feature matching problem for label-free quantification." Bioinformatics **29**(14): 1768-1775.

Magrane, M. and U. Consortium (2011). "UniProt Knowledgebase: a hub of integrated protein data." Database **2011**: bar009.

Martin, J. A., A. Martini, A. Molinari, W. Morgan, W. Ramalingam, J. A. Buckwalter and T. O. McKinley (2012). "Mitochondrial electron transport and

glycolysis are coupled in articular cartilage." Osteoarthritis and Cartilage **20**(4): 323-329.

Matsumoto, E., T. Furumatsu, T. Kanazawa, M. Tamura and T. Ozaki (2012). "ROCK inhibitor prevents the dedifferentiation of human articular chondrocytes." Biochemical and Biophysical Research Communications **420**(1): 124-129.

McLane, L., P. Chang, A. Granqvist, H. Boehm, A. Kramer, J. Scrimgeour and J. Curtis (2013). "Spatial organization and mechanical properties of the pericellular matrix on chondrocytes." Biophysical Journal **104**(5): 986-996.

Michalski, A., E. Damoc, J.-P. Hauschild, O. Lange, A. Wieghaus, A. Makarov, N. Nagaraj, J. Cox, M. Mann and S. Horning (2011). "Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer." Molecular & Cellular Proteomics **10**(9): M111.011015.

Millioni, R., S. Tolin, L. Puricelli, S. Sbrignadello, G. Fadini, P. Tessari and G. Arrigoni (2011). "High Abundance Proteins Depletion vs Low Abundance Proteins Enrichment: Comparison of Methods to Reduce the Plasma Proteome Complexity." PLoS ONE **6**(5): e19603.

Monemdjou, R., F. Vasheghani, H. Fahmi, G. Perez, M. Blati, N. Taniguchi, M. Lotz, R. St-Arnaud, J.-P. Pelletier, J. Martel-Pelletier, F. Beier and M. Kapoor (2012). "Association of cartilage-specific deletion of peroxisome proliferator-activated receptor γ with abnormal endochondral ossification and impaired cartilage growth and development in a murine model." Arthritis and Rheumatism **64**(5): 1551-1561.

Nishida, Y., C. B. Knudson and W. Knudson (2003). "Extracellular matrix recovery by human articular chondrocytes after treatment with hyaluronan hexasaccharides or Streptomyces hyaluronidase." Modern Rheumatology **13**(1): 62-68.

Otte, P. (1991). "Basic cell metabolism of articular cartilage. Manometric studies." Zeitschrift für Rheumatologie **50**(5): 304-312.

Pattappa, G., H. Heywood, J. de Bruijn and D. Lee (2011). "The metabolism of human mesenchymal stem cells during proliferation and differentiation." Journal of Cellular Physiology **226**(10): 2562-2570.

Payne, S. (2015). "The utility of protein and mRNA correlation." Trends in Biochemical Sciences **40**(1): 1-3.

Pfander, D. and K. Gelse (2007). "Hypoxia and osteoarthritis: how chondrocytes survive hypoxic environments." Current Opinion in Rheumatology **19**(5): 457-462.

Poleni, P. E., A. Bianchi, S. Etienne, M. Koufany, S. Sebillaud, P. Netter, B. Terlain and J. Y. Jouzeau (2007). "Agonists of peroxisome proliferators-activated receptors (PPAR) alpha, beta/delta or gamma reduce transforming growth factor (TGF)-beta-induced proteoglycans' production in chondrocytes." Osteoarthritis and Cartilage **15**(5): 493-505.

Ritty, T., R. Roth and J. Heuser (2003). "Tendon cell array isolation reveals a previously unknown fibrillin-2-containing macromolecular assembly." Structure (London, England : 1993) **11**(9): 1179-1188.

Rockstroh, M., S. Müller, C. Jende, A. Kerzhner, M. von Bergen and J. M. Tomm "Cell fractionation - an important tool for compartment proteomics." Journal of Integrated Omics **1**(1): 135-143.

Ruiz-Romero, C., V. Calamia, J. Mateos, V. Carreira, M. Martínez-Gomariz, M. Fernández and F. Blanco (2009). "Mitochondrial dysregulation of osteoarthritic human articular chondrocytes analyzed by proteomics: a decrease in mitochondrial superoxide dismutase points to a redox imbalance." Molecular & Cellular Proteomics **8**(1): 172-189.

Ruiz-Romero, C., V. Calamia, B. Rocha, J. Mateos, P. Fernández-Puente and F. Blanco (2010). "Hypoxia conditions differentially modulate human normal and osteoarthritic chondrocyte proteomes." Journal of Proteome Research **9**(6): 3035-3045.

Schwämmle, V., I. R. León and O. N. Jensen (2013). "Assessment and improvement of statistical tools for comparative proteomics analysis of sparse data

sets with few experimental replicates." Journal of Proteome Research **12**(9): 3874-3883.

Smyth, G. K. (2005). limma: Linear Models for Microarray Data. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. R. Gentleman, Carey, V.J., Huber, W., Irizarry, R.A., and Dudoit, S. New York, Springer**:** 397-420.

Thiede, B., W. Höhenwarter, A. Krah, J. Mattow, M. Schmid, F. Schmidt and P. Jungblut (2005). "Peptide mass fingerprinting." Methods (San Diego, Calif.) **35**(3): 237-247.

Tian, Q., S. Stepaniants, M. Mao, L. Weng, M. Feetham, M. Doyle, E. Yi, H. Dai, V. Thorsson, J. Eng, D. Goodlett, J. Berger, B. Gunter, P. Linseley, R. Stoughton, R. Aebersold, S. Collins, W. Hanlon and L. Hood (2004). "Integrated Genomic and Proteomic Analyses of Gene Expression in Mammalian Cells." Molecular & Cellular Proteomics **3**(10): 960-969.

Ting, L., M. Cowley, S. L. Hoon, M. Guilhaus, M. Raftery and R. Cavicchioli (2009). "Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling." Molecular & Cellular Proteomics **8**(10): 2227-2242.

Tomita, T., Y. Kakiuchi and P. Tsao (2006). "THR0921, a novel peroxisome proliferator-activated receptor gamma agonist, reduces the severity of collagen-induced arthritis." Arthritis Research & Therapy **8**(1): R7.

Tsuchida, A., M. Beekhuizen, M. Rutgers, G. van Osch, J. Bekkers, A. Bot, B. Geurts, W. Dhert, D. Saris and L. Creemers (2012). "Interleukin-6 is elevated in synovial fluid of patients with focal cartilage defects and stimulates cartilage matrix production in an in vitro regeneration model." Arthritis Research & Therapy **14**(6): R262.

Vasheghani, F., R. Monemdjou, H. Fahmi, Y. Zhang, G. Perez, M. Blati, R. St-Arnaud, J.-P. Pelletier, F. Beier, J. Martel-Pelletier and M. Kapoor (2013). "Adult cartilage-specific peroxisome proliferator-activated receptor gamma knockout

mice exhibit the spontaneous osteoarthritis phenotype." The American Journal of Pathology **182**(4): 1099-1106.

Vogel, C., R. d. S. Abreu, D. Ko, S.-Y. Le, B. Shapiro, S. Burns, D. Sandhu, D. Boutz, E. Marcotte and L. Penalva (2010). "Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line." Molecular Systems Biology **6**(1): 400.

Wang, H., Q. Wang, U. Pape, B. Shen, J. Huang, B. Wu and X. Li (2010). "Systematic investigation of global coordination among mRNA and protein in cellular society." BMC Genomics **11**: 364.

Wang, M.-C., G. Pins and F. Silver (1995). "Preparation of fibrin glue: the effects of calcium chloride and sodium chloride." Materials Science and Engineering: C **3**(2): 131-135.

Wanner, J., R. Subbaiah, Y. Skomorovska-Prokvolit, Y. Shishani, E. Boilard, S. Mohan, R. Gillespie, M. Miyagi and R. Gobezie (2013). "Proteomic profiling and functional characterization of early and late shoulder osteoarthritis." Arthritis Research & Therapy **15**(6): R180.

Wilson, R., J. Whitelock and J. Bateman (2009). "Proteomics makes progress in cartilage and arthritis research." Matrix Biology **28**(3): 121-128.

Wilusz, R. E., J. Sanchez-Adams and F. Guilak (2014). "The structure and function of the pericellular matrix of articular cartilage." Matrix Biology **39**(0): 25-32.

Woods, A. and F. Beier (2006). "RhoA/ROCK signaling regulates chondrogenesis in a context-dependent manner." The Journal of Biological Chemistry **281**(19): 13134-13140.

Wu, J., S. Luo, H. Jiang and H. Li (2005). "Mammalian CHORD-containing protein 1 is a novel heat shock protein 90-interacting protein." FEBS Letters **579**(2): 421-426.

Yin, W., J.-I. Park and R. Loeser (2009). "Oxidative stress inhibits insulin-like growth factor-I induction of chondrocyte proteoglycan synthesis through differential regulation of phosphatidylinositol 3-Kinase-Akt and MEK-ERK

MAPK signaling pathways." The Journal of Biological Chemistry **284**(46): 31972-31981.

Yoshida, S., S. Tsutsumi, G. Muhlebach, C. Sourbier, M.-J. Lee, S. Lee, E. Vartholomaiou, M. Tatokoro, K. Beebe, N. Miyajima, R. Mohney, Y. Chen, H. Hasumi, W. Xu, H. Fukushima, K. Nakamura, F. Koga, K. Kihara, J. Trepel, D. Picard and L. Neckers (2013). "Molecular chaperone TRAP1 regulates a metabolic switch between mitochondrial respiration and aerobic glycolysis." Proceedings of the National Academy of Sciences **110**(17): E1604-E1612.

Zhang, Z., W. Jin, J. Beckett, T. Otto and B. Moed (2011). "A proteomic approach for identification and localization of the pericellular components of chondrocytes." Histochemistry and Cell Biology **136**(2): 153-162.

7: Exploration of integration strategies for proteomic and transcriptomic data sets

Abstract

Omics data integration poses considerable technical and conceptual challenges. The principle motivation is a deeper understanding of the mechanistic relationships between the components comprising a living system and describing that through mathematical or relational models.

Conflicting pathway analysis and upstream regulators across transcriptomic and proteomic data sets prompted exploration of tools for omics integration to define consensus pathways and regulators. It was proposed that rationalization of the key regulators in de- and re-differentiation would yield a targetable mechanistic network. Data from two transcriptomic and one proteomic study were integrated on three levels: i) union of discrete elements, ii) by functional annotations and, iii) by mechanistic networks derived from common upstream regulators.

Integration of mechanistic networks defined a protein-protein interaction network centred on the predicted reciprocal mediation by TGFβ and PDGF-BB. Common downstream targets and intermediate regulators were identified for both cartilage and tendon from mechanistic networks including CCN2/connective tissue growth factor and SMAD7 in TGF-β1-regulated networks.

To fully utilize the multi-level omics data available in this thesis data imputation and quantitative integration methods should be explored. Simple statistical strategies for imputation of missing data may be appropriate, whilst multivariate projection-based approaches could be employed to integrate quantitative data from gene and protein profiling studies.

7.1: Introduction

7.1.1: Motivations for multi-omics data integration

Data discovery and data exploitation are the two challenges that comprise data integration and are central to the principal goals of systems biology research, i) defining the components of the living systems, and ii) understanding the dysfunction of the system arising from the interaction of these components (Gomez-Cabrero, Abugessaisa et al. 2014). The hallmark responses to changes in conditions are captured within the gene response and this response may then be subsequently profiled at multiple levels in the biological hierarchy (proteome, metabolome, methylome). Data integration assumes that each of the functional levels of the biological hierarchy are inter-related and that consideration of all elements of these levels as a whole system will decipher the complexity of living organisms. By integrating these levels profiling becomes more comprehensive and a more robust understanding of the active processes may become reliably described (Sass, Buettner et al. 2013).

As presented in **Chapter 6** the correlation of effect sizes between transcriptomic and proteomic studies is only moderate as a consequence of a number of intermediary events for which information is unavailable, e.g. half-life and kinetic data, post-transcriptional modifications influencing activation status, and temporal information. Furthermore, conflicting pathway prediction and inconsistencies in functional annotation of data sets prompts further exploration of the data sets to determine a consensus systems account of the responses of chondrocytes and tenocytes to culture conditions. Integration of transcriptomic and proteomic data

can provide another level of data analysis and identify changes in the samples not apparent from independent analysis alone (Haider and Pal 2013).

Being able to integrate data from these different levels would encourage a more robust understanding of the key regulatory mechanisms at play in the de- and re-differentiation transitions associated with organotypic cultures for culture and tendon. It was hypothesised that use of omics integration strategies would facilitate the development of unified protein-protein network for de- and re-differentiation establishing a testable model for future analysis. It was proposed that these techniques would generate a consensus understanding of the novel data sets generated in this thesis and rationalise conflicts in predicted pathways and core regulators.

7.1.2: Challenges of integration

"*Know less, faster*". Prof. Harvey Blanch

As the cost of per unit of measurement using omics technologies has decreases the data generated has increased. This has led to numerous challenges not least of which is the outpacing of resource allocation for processing and integration by the data generated (Palsson and Zengler 2010). Although the hurdles to omics integration are significant, a number of web-based (e.g. 3Omics, (Kuo, Tian et al. 2013)) and open-source software tools (*R* packages, Cytoscape) have recently become available to support this type of investigation, in addition to public repositories and consortia-based projects (Cabrero, Abugessaisa et al. 2014). There

are a number of general approaches that may be used for integration, each varying in complexity.

Union of the data elements represents the most straightforward approach and creates a reference data set, however this considers only absolute matches of elements (or features) and does not represent overlapping functional descriptors of the data or deal with missing data. The second approach deals with extracting common functional annotations for both data sets, for example, enrichment analysis for gene ontology biological process terms or signaling pathways. Topological network approaches (see **Chapters 4** and **5**), meta-analysis merging of individual domain datasets and correlations, estimation of missing data, multiple regression analysis, clustering approaches and dynamic modeling represent more complex techniques to omics integration (Haider and Pal 2013).

The methods employed to integrate the data depend on several factors, including: a) an understanding of the source of the data and its limitations; b) an awareness of the differences in expression between the two domains and why they may not correlate; c) a clear objective for the analysis; d) and whether the combined model may be extrapolated or represents a collective behaviour of a group of cells within a tissue and not the system in general (Haider and Pal 2013).

Core to the consideration of high-throughput, multi-dimensional data is the question of whether the data arising is derived from direct measurement of the biological features or as a consequence of the experimental condition, and/or bias and artefacts. Unsupervised analysis, such as clustering, model-based and projection-based methods, are exploratory statistical approaches that allow

reduction of dimensionality, and noise, and allows data to be visualized in a smaller subspace by graphical representations (Yao, Coquery et al. 2012).

7.1.3: Multivariate data analysis methods

Researchers have used a number of multi-variate analysis approaches to integrate highly dimensional data from diverse data sources. These 'projection-based' methods, principal component analysis (PCA) being the classical example, integrate the data through the projection of each data set into smaller subspace to maximize the covariance between data sets (Günther, Shin et al. 2014). This allows the key players to be identified. In addition to PCA projection-based methods include independent PCA (Yao, Coquery et al. 2012), partial least squares regression (PLS) (Lê Cao, Boitard et al. 2011), canonical correlation analysis (CCA) (González, Déjean et al. 2008). Unlike PCA these methods have the capacity to integrate two types of data sets. They are also widely implemented in *R* packages, e.g. `mixOmics` (Lê Cao, González et al. 2009).

Using a modification of co-inertia analysis Meng, *et al* (2014) integrated transcriptomic and proteomic data from multiple cancerous cell lines from a variety of tissues (`omicade4` *R* package) (Meng, Kuster et al. 2014). This strategy had the advantage of not requiring mapping or filtering the different data sets to intersect on common features across more than two data sets. The methodology provides graphical representation of the relationships between the data sets, but requires the multiple data sets to arise from the same individuals.

7.1.4: Project Aims

Using the transcriptomic and proteomic data sets arising from **Chapters 2**, **3** and **6** several integration strategies were explored, i) union of data set elements, ii)

extraction of common feature annotations, iii) derivation of common upstream regulators and development of mechanistic networks. Here the intention is to define common regulatory mechanisms for de- and re-differentiation and consider how these are associated with current understanding of disease mechanisms. Furthermore, by defining the common regulatory mechanisms this may focus efforts in the research and development of organo-typic culture systems.

7.2: Methods

7.2.1 Extraction of common features and functional annotations

Union of discrete elements

All differential expressed genes and proteins were converted to unique Entrez gene identifiers and common elements were extracted for each pairwise comparison. Overlaps between each data set are presented as Euler plots.

RAMONA

To integrate multi-level data in terms of gene ontology functional annotation the RAMONA web application was used (http://icb.helmholtz-muenchen.de/ramona) (Sass, Buettner et al. 2014) an implementation of the multi-level ontology analysis (MONA) method (Sass, Buettner et al. 2013). Briefly, Entrez identifiers for the rat arising from differential expression analysis for either transcriptomic or proteomic data was used as the input to a model-based Bayesian approach to infer term probabilities (P). Different pair-wise comparisons were considered in turn. The output from RAMONA contained some redundant annotations. In order to limit descriptors to unique terms REVIGO was employed (as before, **Chapter 2**) to filter on semantic similarity. Gene ontology terms with a term probability of $P > 0.5$ were used as the input to REVIGO. Presented terms had a probability of > 0.5 (RAMONA) and threshold dispensability score < 0.3 (REVIGO). All results are available in **SD7.**

7.2.2: Defining common upstream regulators and developing mechanistic networks

Topological network approaches implemented within Ingenuity® Pathway Analysis were used as previously described (**2.2.5**) on individual data sets to ascertain the common upstream regulators, create mechanistic networks and annotate canonical pathways. Analysis for common regulators was run on the Ingenuity® knowledge base issued November 2014. Mechanistic networks were derived by connecting sets of upstream regulators within the IPA knowledge base that could elicit the gene or protein expression profiles provided from the data sets. Known downstream targets of the regulators within the differential expression data set could be added to the network. Nodes (genes) were added to mechanistic networks to provide a cell-specific context as defined in the text, for example TNMD and COL2A1 for tendon and cartilage respectively.

7.3: Results

7.3.1: Union of discrete elements

Using only universal Entrez annotations and removing duplicate entries there were few data set elements that were found to be common to all three data sets, **Figure 7.1 – 7.4.** The greatest overlap between transcriptomic and proteomic data sets was for the cartilage to monolayer chondrocyte comparison (130 features) and the fewest were found for the comparison between monolayer tenocytes and fibrin constructs (41 features). Despite the Illumina transcriptomic data arising from a different experimental protocol the number of features overlapping with the proteomic data was comparable between the Affymetrix and Illumina transcriptomic data sets. Given the low numbers of features that were consistently identified across all data sets union of unique elements was not considered an adequate method for integrating multi-omics data sets.

7.3.2: Integrated gene ontology functional annotations using RAMONA

CHONDROCYTES

Dedifferentiation in chondrocytes, following integration of biological process gene ontology terms across data sets, was associated with the terms: 'protein transport', 'ossification', 'I-kappaB kinase/NF-kappaB signaling', 'Wnt-signaling pathway', 'oxidation-reduction process' and included development-associated terms 'regulation of cell morphogenesis', 'cardiovascular system development' and 'anatomical structure morphogenesis', **Figure 7.1**. Enriched KEGG pathways using gene ontology across all datasets were: 'RNA transport', 'Wnt-signaling pathway' and 'osteoclast differentiation'.

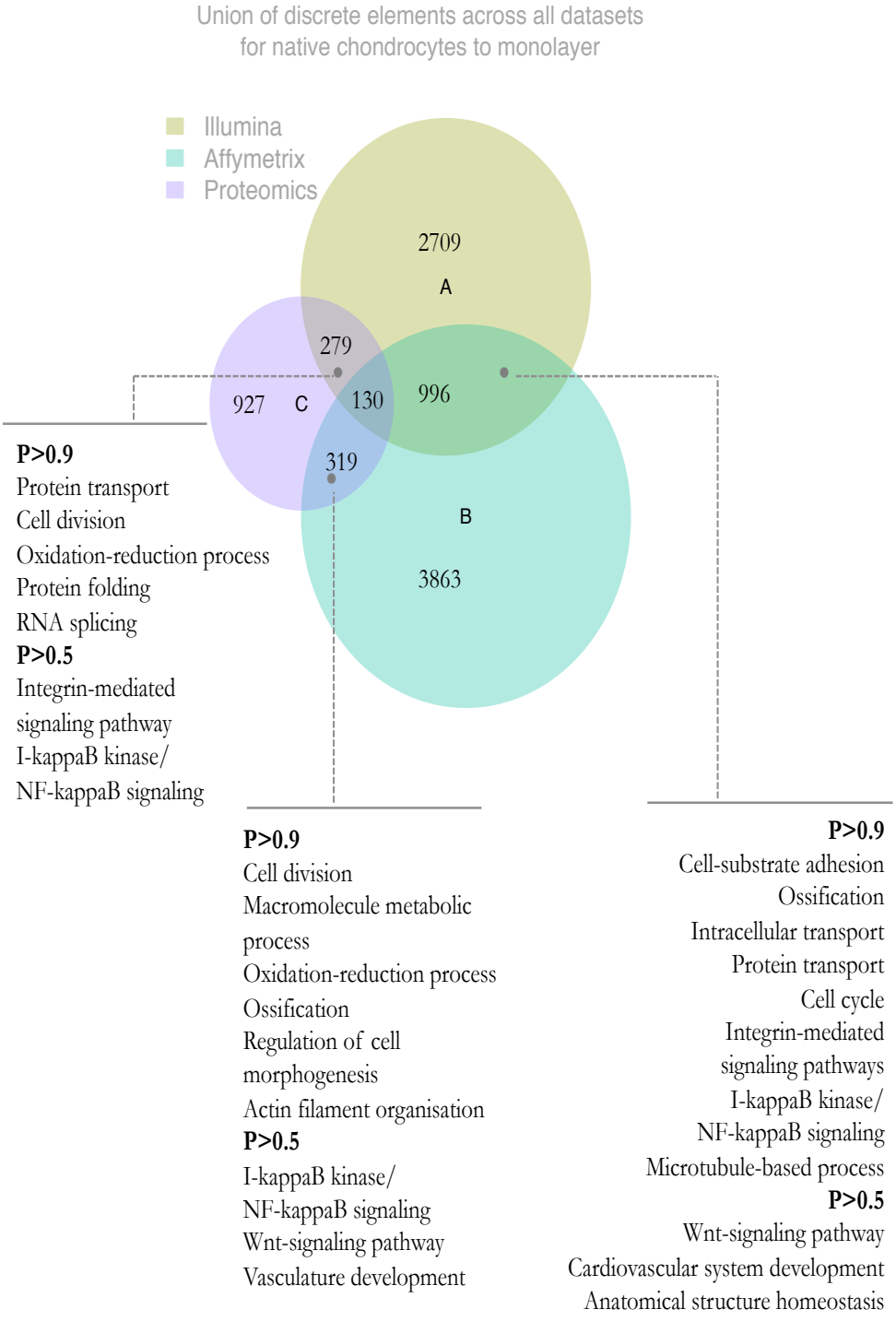
For the re-differentiation transition (monolayer to alginate beads), **Figure 7.2**, functional annotation across all studies was associated with mitochondrial associated terms - 'response to oxidative stress', 'oxidation-reduction process', 'proton transport', 'electron transport chain'. Other terms included: 'muscle structure development', 'collagen fibril organisation' and 'integrin-mediated signaling pathway'. Enriched KEGG pathways related to the canonical pathways 'proteosome', 'RNA transport', and 'oxidative phosphorylation'.

TENOCYTES

In the dedifferentiation transition from native tenocytes to monolayer functional annotation across data sets was associated with: 'oxidation-reduction process', 'ossification', 'muscle structure development', 'actin filament-based process' and 'connective tissue development', **Figure 7.3**. KEGG canonical pathways enriched in this analysis related to 'RNA transport', 'ribosome', 'Wnt-signaling pathway' and 'axon guidance'.

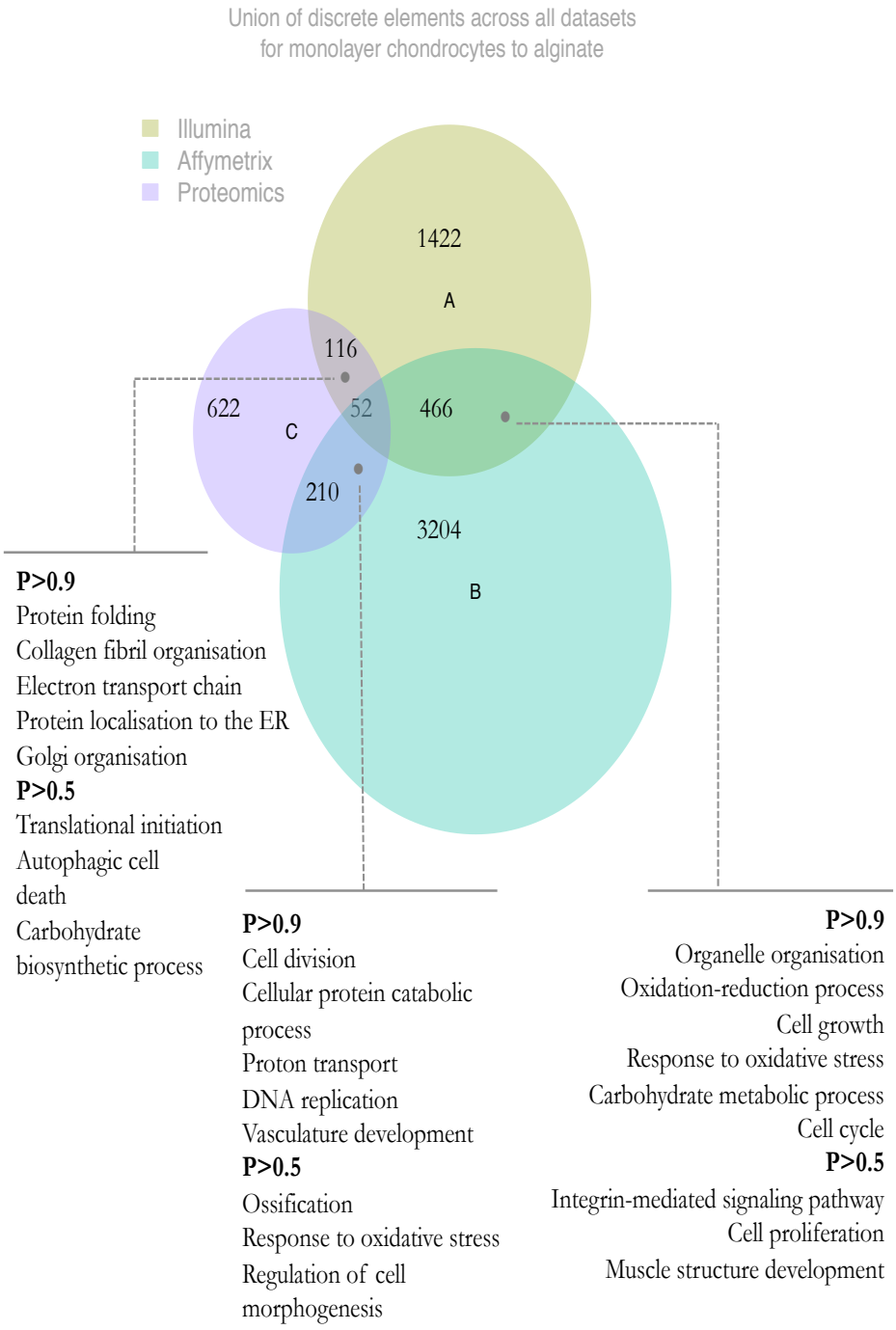
The transition to tenogenic fibrin constructs from monolayer was annotated with 'response to wounding', 'ECM organization', 'oxidation-reduction process' and 'cellular response to glucose starvation', **Figure 7.4**. Additionally, 'phospholipid biosynthetic process' and 'cellular response to cytokine stimulus' were notable. Enriched KEGG pathways were comparable to those found for chondrocyte redifferentiation – 'aminoacyl-tRNA biosynthesis', 'proteosome', 'Parkinson's disease' and 'oxidative phosphorylation'.

These unified annotations were consistent with findings presented for each data set independently. Data contributing to cross data set functional annotation analysis is found in **SD7**.



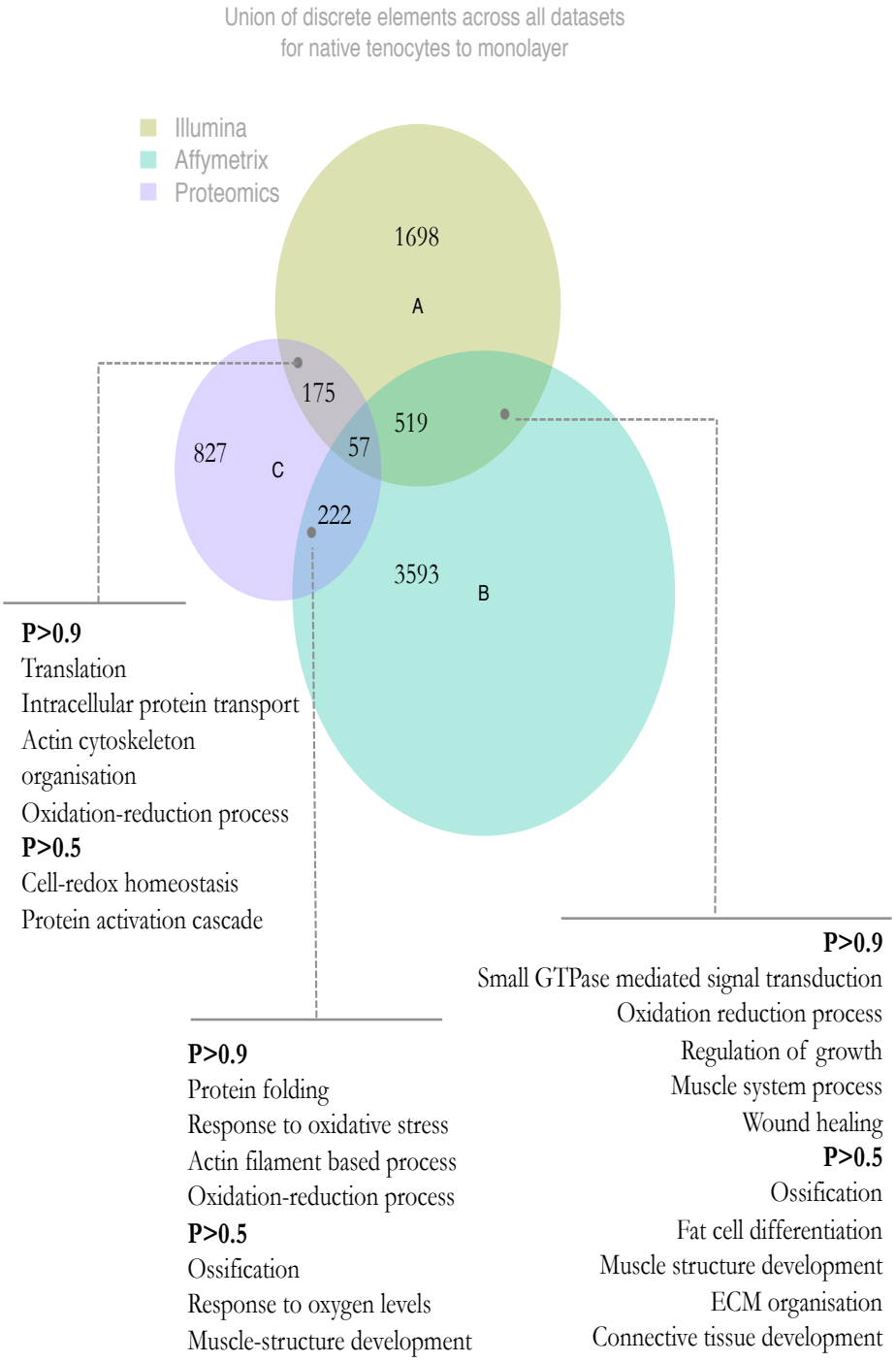
Commonly enriched KEGG pathways using RAMONA: RNA transport; Wnt-signaling pathway, osteoclast differentiation

Figure 7.1: Union of Entrez identifiers from transcriptomic and proteomic analysis for the transition from **native chondrocytes to monolayer chondrocytes**. Euler plot shows the number of Entrez gene identifiers shared between data sets once duplicates have been removed. Term probabilities (P) as defined by RAMONA are divided into those >0.5 and those >0.9.



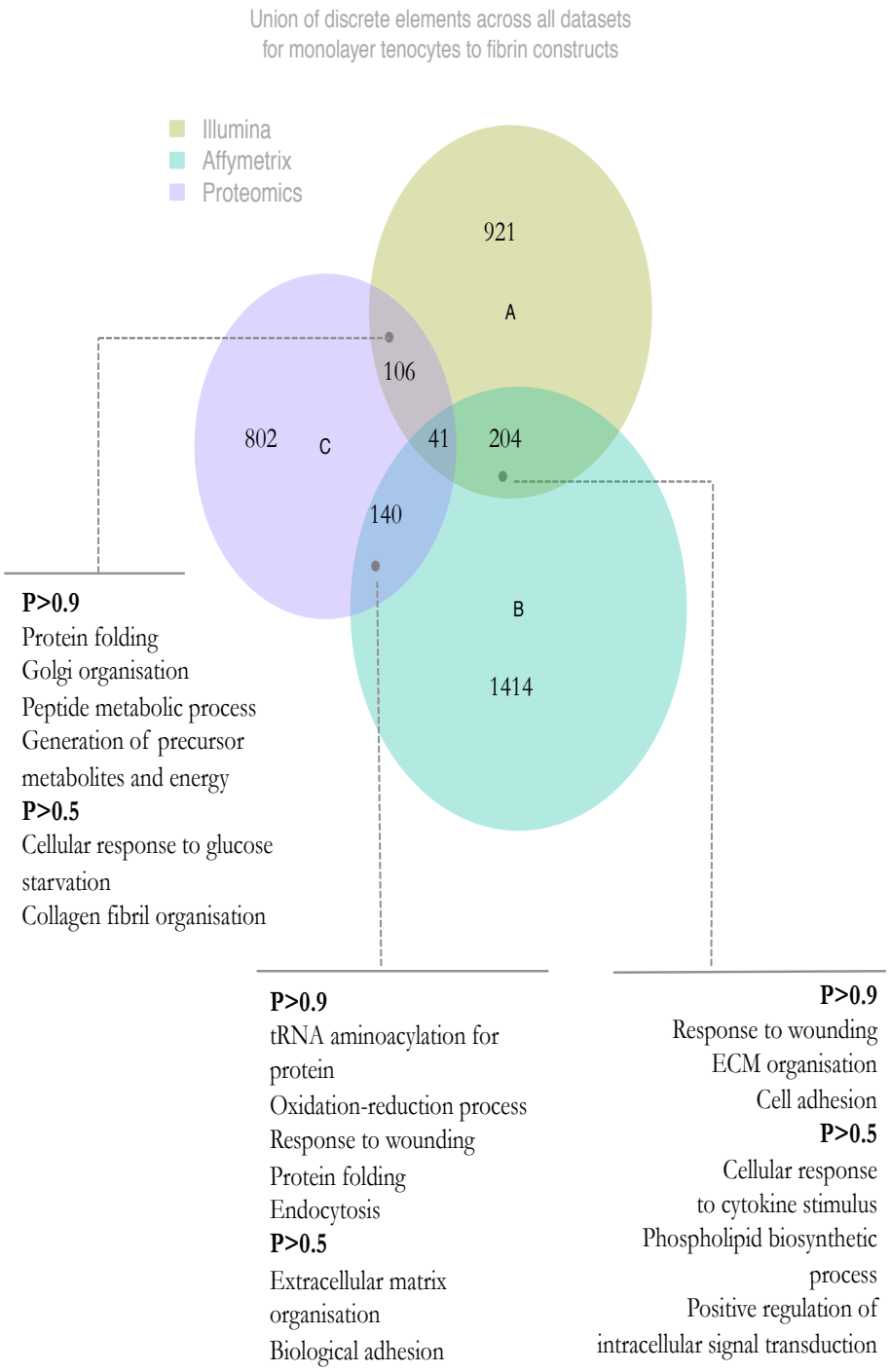
Commonly enriched KEGG pathways using RAMONA: Proteosome, Protein processing in the endoplasmic reticulum; lysosome, RNA transport, oxidative phosphorylation and valine, leucine, and isoleucine degradation. No significant enrichment of KEGG pathways was found for cross-transcriptome analysis using RAMONA.

Figure 7.2: Union of Entrez identifiers from transcriptomic and proteomic analysis for the transition from **monolayer chondrocytes to alginate cultures**. Euler plot shows the number of Entrez gene identifiers shared between data sets once duplicates have been removed. Term probabilities (P) as defined by RAMONA are divided into those >0.5 and those >0.9.



Commonly enriched KEGG pathways using RAMONA: RNA transport, ribosome, Wnt-signaling pathway, axon guidance, lysosome.

Figure 7.3: Union of Entrez identifiers from transcriptomic and proteomic analysis for the transition from **native tenocytes to monolayer cultures**. Euler plot shows the number of Entrez gene identifiers shared between data sets once duplicates have been removed. Term probabilities (P) as defined by RAMONA are divided into those >0.5 and those >0.9.



Commonly enriched KEGG pathways using RAMONA: Protein processing in endoplasmic reticulum; Aminoacyl-tRNA biosynthesis; proteosome; Parkinson's disease; oxidative phosphorylation; insulin signaling pathway. No significant enrichment of KEGG pathways was found for cross-transcriptome analysis using RAMONA.

Figure 7.4: Union of Entrez identifiers from transcriptomic and proteomic analysis for the transition from **monolayer tenocytes to fibrin cultures**. Euler plot shows the number of Entrez gene identifiers shared between data sets once duplicates have been removed. Term probabilities (P) as defined by RAMONA are divided into those >0.5 and those >0.9.

7.3.3 Common upstream regulators define de- and re-differentiation mechanistic networks

Integration by intersection of discrete elements or functional ontologies fails to provide a mechanistic understanding of the data or make use of quantitative data. By defining upstream regulators common to de- or re-differentiation conditions for chondrocytes or tenocytes it may be possible to define a targetable mechanistic network. Using Ingenuity® Pathway Analysis the upstream master regulators, consistent with differential expression profiles, were predicted. Regulators that were consistently changed across a transition and cell-type were considered. Mechanistic networks were created for common regulators. These networks were derived from known protein-protein interactions and evidence from the curated IPA knowledge base.

Dedifferentiation is associated with a TGF-β1 network

In general, for the transition to monolayer from native cells TGF-β1 was predicted to be a common upstream regulator (**Table 7.1**). Mechanistic networks derived from differential expression lists predicted TGF-β1 to be inhibited in the native context i.e. genes or proteins found to have higher expression in native cartilage or tendon were consistent with a predicted inhibition of TGF-β1 regulated networks. Activation status was variable across data sets with low z-score in the Affymetrix chondrocyte data set, however, TGF-β1 had significantly lower expression in native chondrocytes and tenocytes compared to monolayer consistent with IPA prediction.

SMAD7 and CCN2 are common downstream targets

For both chondrocyte and tenocyte dedifferentiation TGF-β1 mechanistic networks SMAD7 was predicted as an intermediate regulator and down-stream target of TGF-β1. SMAD7 was down-regulated in native cartilage and chondrocytes in both

Illumina and Affymetrix gene expression data sets and in native tenocytes in the Affymetrix data set. Across transcriptomic and proteome profiling studies the CCN family member CCN2/connective tissue growth factor (CTGF) also showed lower expression in native cells relative to monolayer and was a down-stream target of TGF-β1 and SMAD7, which was predicted to inhibit CCN2 expression.

A dedifferentiation mechanistic network was developed comprising the predicted upstream regulators, consistent with the provided expression profiles, and their downstream targets. Quantitative expression profiles were overlain for each data set and comparison, **Figures 7.5-7.7**. A unified dedifferentiation model was prepared based upon the common intermediate regulators and down-stream effectors frequently found in the differential expression analysis, including CCN2 and THY-1, **Figures 7.8-7.9**. A global protein-protein interaction network (STRING) of this proposed dedifferentiation mechanistic network was found to be highly enriched for interaction and functional annotations related to cartilage development and condensation, **Figure 7.10**, consistent with predictions from the IPA knowledge base.

Cell type	Data source	Upstream regulator	Activation z-score	Overlap p-value	DE genes available
Chondrocytes	Illumina	TP53	-3.68	9e-32	435 (14)
		TGFB1	**-2.59**	2.04e-26	679 (23)
		MYC	4.08	1.9e-18	515 (19)
Chondrocytes	Affymetrix	**TGFB1**	**-0.52**	3.2e-25	726 (19)
		TNF	5.48	1.14e-24	663 (17)
		PDGFBB	3.64	2.4e-23	477 (18)
Chondrocytes	Proteomics	MYC	-4.71	7.2e-61	393 (18)
		NFE2L2	-6.56	1.1e-30	198 (8)
		TGFB1	**-4.2**	1.5e-23	384 (18)
Tenocytes	Illumina	**TGFB1**	**-1.04**	1e-23	319 (18)
		HRAS	-0.285	1.2e-20	325 (19)
		KRAS	-2.27	1.3e-16	306 (17)
Tenocytes	Affymetrix	TNF	6.63	3.7e-50	692 (17)
		PDGFBB	3.86	4.7e-39	549 (23)
		TGFB1	**-2.2**	3.9e-38	685 (20)
Tenocytes	Proteomics	MYC	-4.35	6.7e-62	399 (13)
		NFE2L2	-6.8	1.04e-25	392 (17)
		TGFB1	**-6.8**	2.1e-19	437 (19)

Table 7.1: Top upstream regulators predicted using Ingenuity Pathway Analysis knowledge base from differential expression/abundance for the **native to monolayer** comparison across three data sources and two cell types. Small molecules and other chemicals predicted by IPA are not included. The activation **z-score** indicates the predicted activation status of the mechanistic network (negative values indicating inhibition). Only genes with available mechanistic networks are shown. **DE genes available** indicates those genes in the differential lists that are known down-stream targets of a number of master regulators (defined in parentheses). TGFB1 is consistently predicted to to be inhibited for gene profiles more highly expressed in native tendon or cartilage, i.e. dedifferentiation is associated with active TGFB1 regulation.

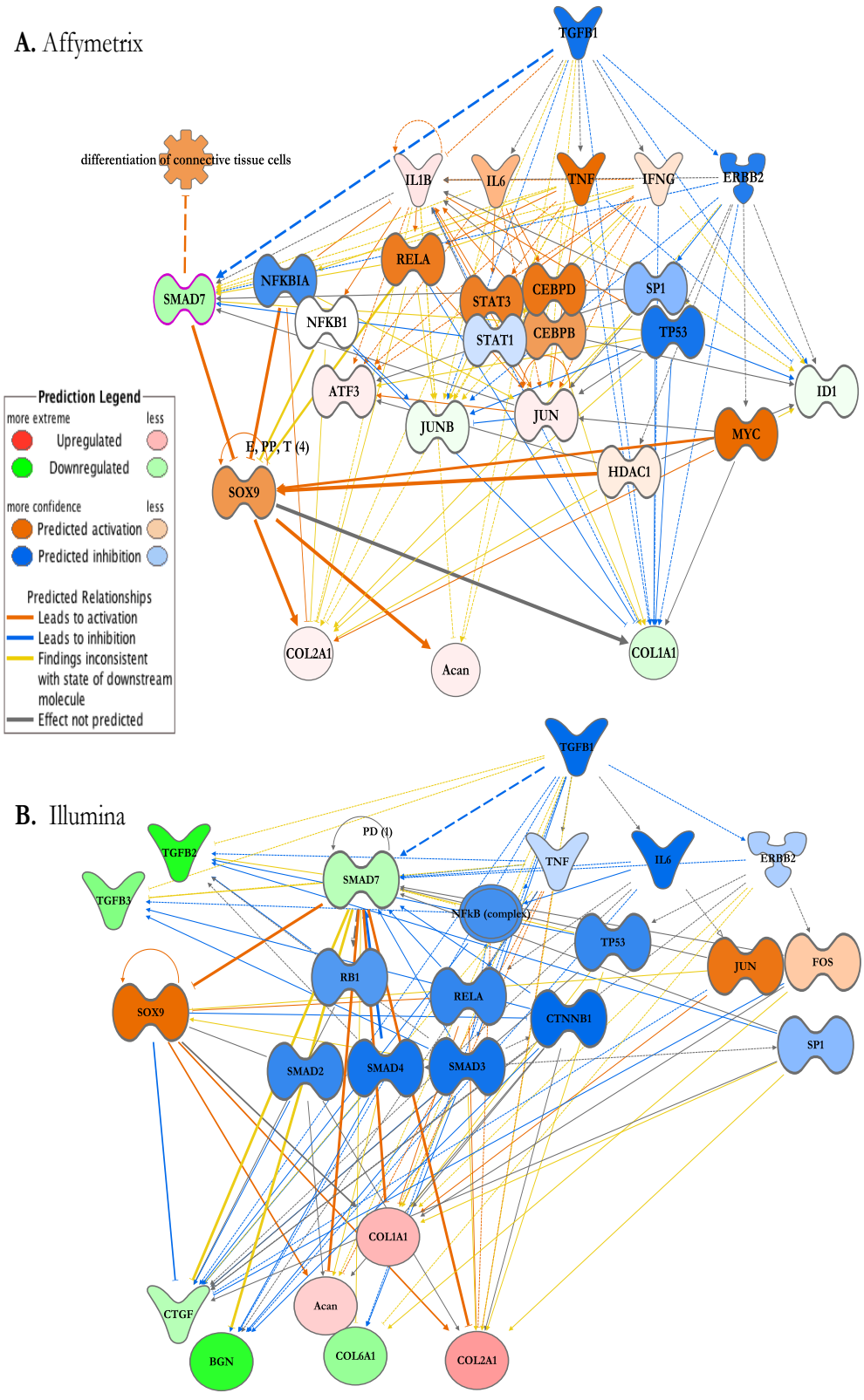


Figure 7.5: Ingenuity mechanistic network formed from Affymetrix (**A**) or Illumina (**B**) gene expression data for native chondrocyte to monolayer transition derived from TGF-β1 network. SMAD7 was found to be a common downstream regulator within the TGF-β1 mechanistic network – expression was significantly lower in native chondrocytes relative to monolayer (figure legend). Only those genes with direct up- or down-stream associations with SMAD7 are shown and annotated with gene symbols. Added nodes: SOX9.

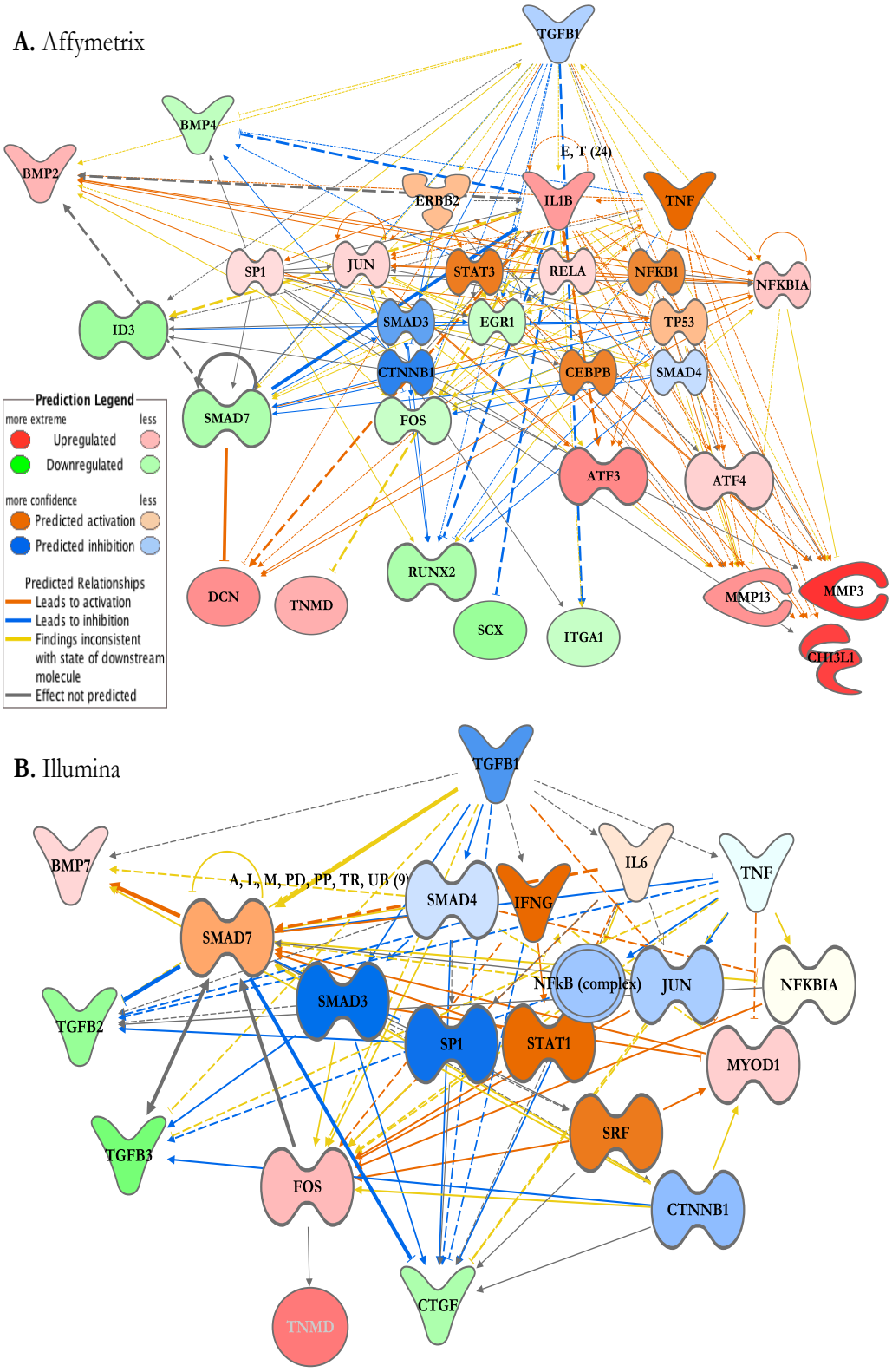


Figure 7.6: Ingenuity mechanistic network formed from Affymetrix (**A**) or Illumina (**B**) data for native tenocyte to monolayer transition derived from TGF-β1 network. In native tenocytes IL-1B was more highly expressed than in monolayer – it was predicted to have an inhibitor effect (figure legend) on SMAD7, SCX and RUNX2, which all had lower expression in native tenocytes than in monolayer. SMAD7 was not differentially expressed in Illumina data – Ingenuity predicted activation which contradicted the findings in the other data sets where SMAD7 was down-regulated in native tissue. Added node: TNMD (**B**.).

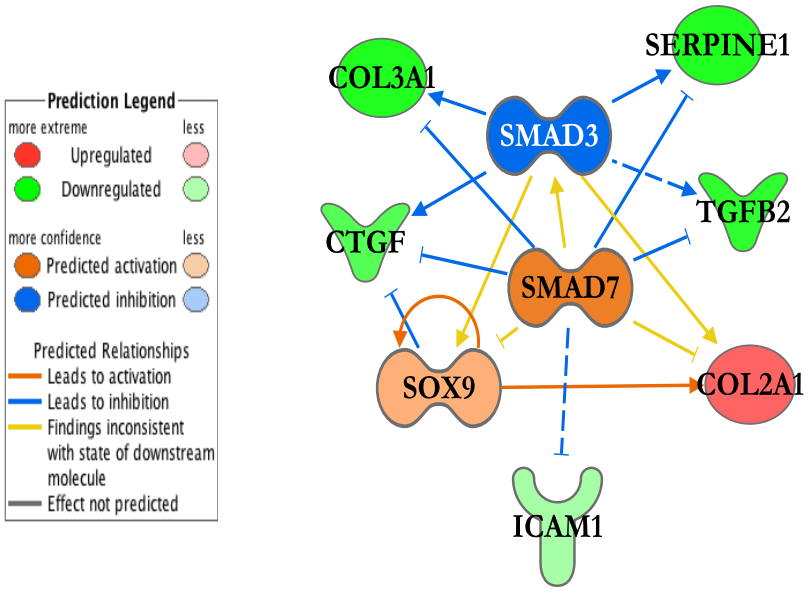


Figure 7.7A: Detail of mechanistic network for **proteomics** data for native chondrocytes to monolayer comparisons. Inhibitor of TGF beta and BMP signalling, SMAD7, was a common intermediate regulator in the TGFB1 mechanistic network and is shown to have inhibitory effects on CNN2/CTGF, SERPINE1 and COL3A1 when active (figure legend). Nodes in green indicate proteins with lower abundance in native chondrocytes. Gene symbols are used for node annotation.

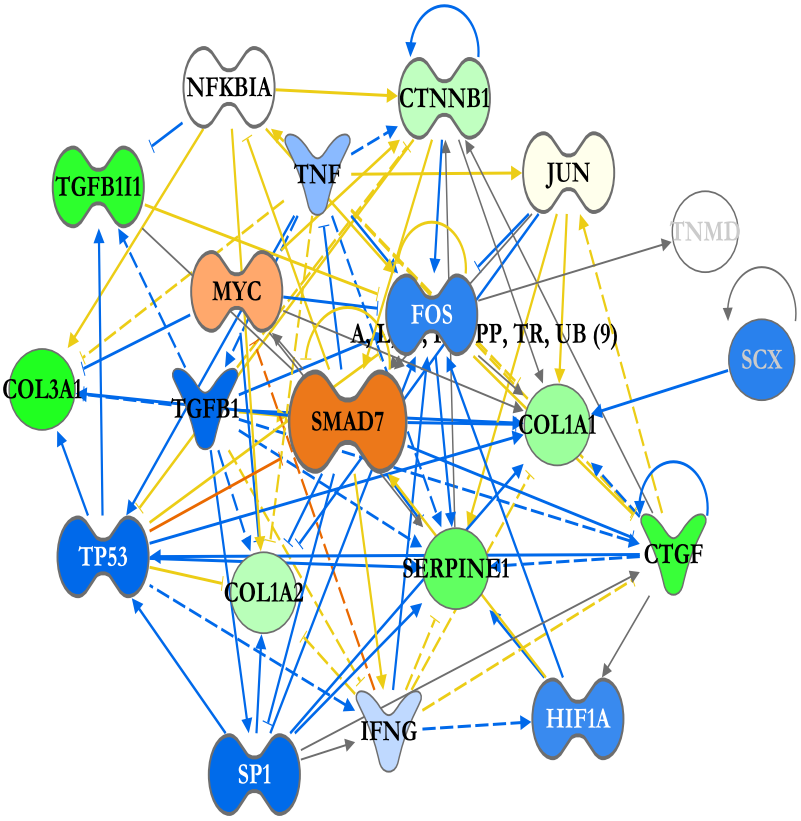


Figure 7.7B: Detail of mechanistic network for **proteomics** data for native tenocytes to monolayer cultures comparisons. JUN was a common intermediate regulator for differentially abundant proteins. Added nodes: TNMD and SCX. In both proteomics analysis SMAD7 is predicted as activated (figure legend). Nodes in green represent proteins with lower expression in native tenocytes relative to monolayer. Gene symbols are used for node annotation.

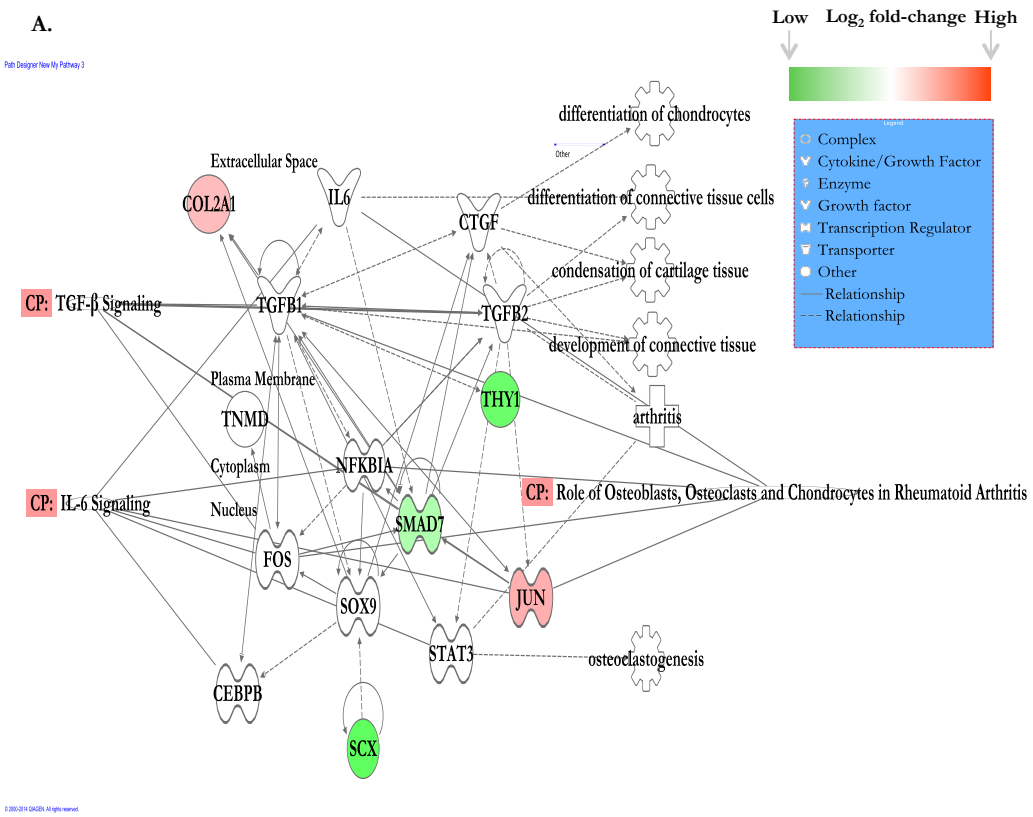


Figure 7.8A: Model network to unify findings from **Figures 7.5-7.7**. Network derived from prediction of Tgf-β1 as a core upstream regulator associated with native cartilage transition to monolayer. Predicted intermediate regulators and tendon markers are overlain with differential expression values from **A**) Affymetrix. Reduction in expression of Smad7 and Thy-1 in native cells is consistent feature of dedifferentiation in transcriptomic data (figure legend). **CP** – canonical pathways; Functional annotation p<0.001.

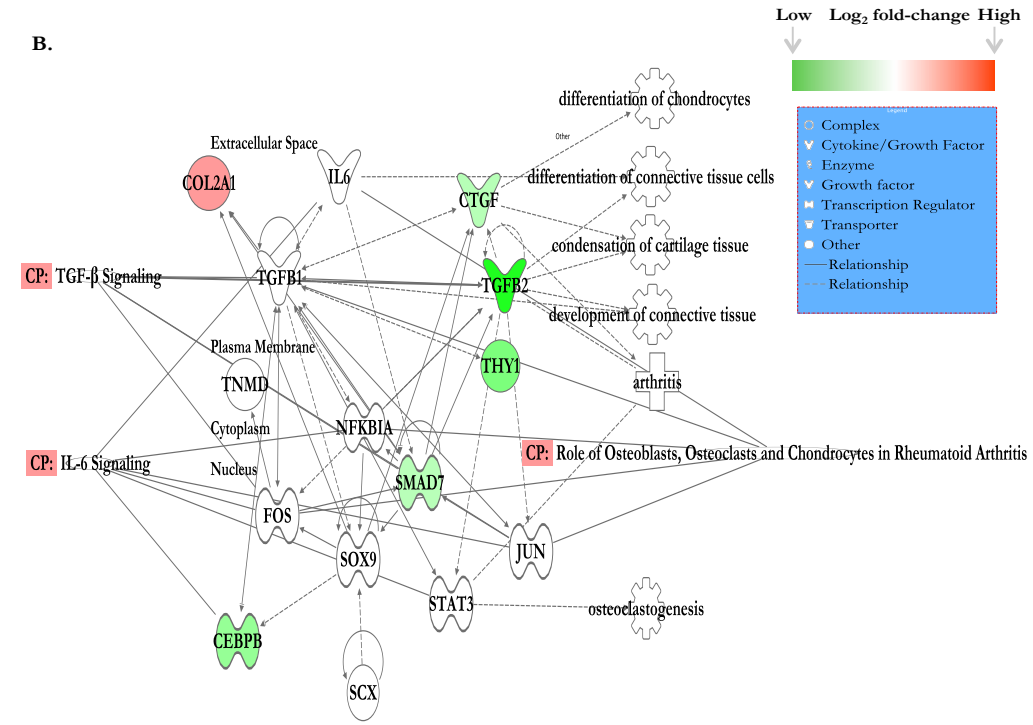


Figure 7.8B: Model network to unify findings from **Figures 7.5-7.7**. Network derived from prediction of Tgf-β1 as a core upstream regulator associated with native cartilage transition to monolayer. Predicted intermediate regulators and tendon markers are overlain with differential expression values from **B**) Illumina. Reduction in expression of Smad7 and Thy-1 in native cells is consistent feature of dedifferentiation in transcriptomic data (figure legend). **CP** – canonical pathways; Functional annotation p<0.001.

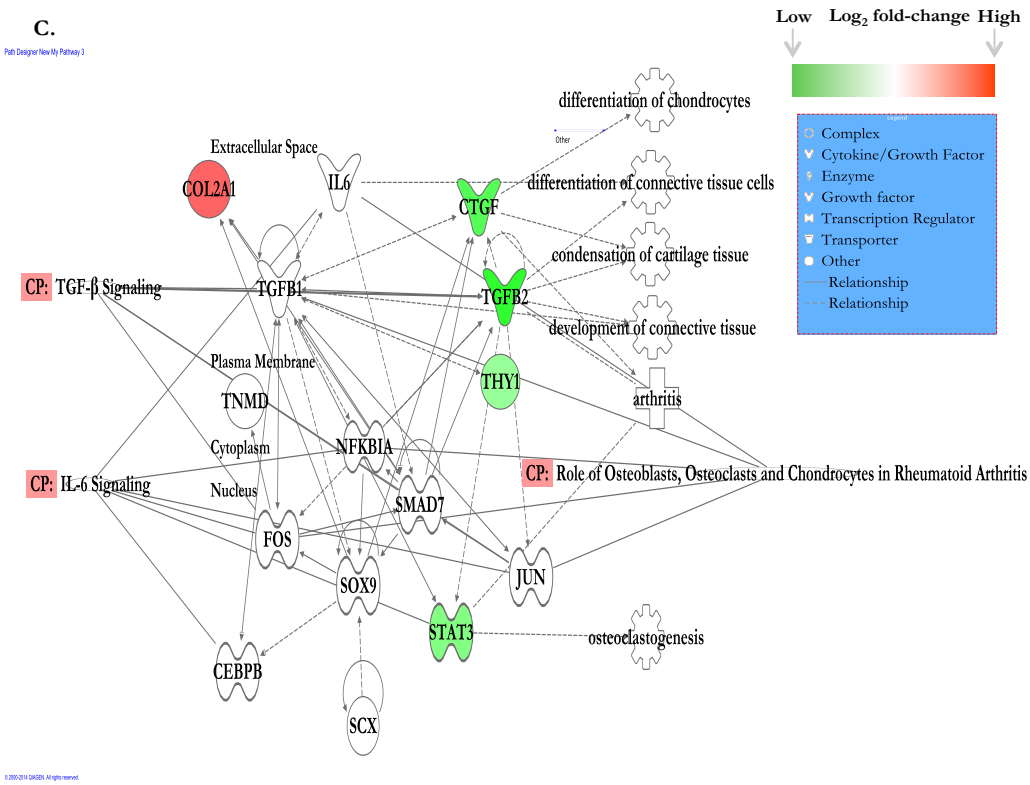


Figure 7.8C: Model network to unify findings from **Figures 7.5-7.7**. Network derived from prediction of Tgf-β1 as a core upstream regulator associated with native cartilage transition to monolayer. Predicted intermediate regulators and tendon markers are overlain with differential abundance values from **C**) proteomic relative quantification study. Reduction in abundance of Ctgf, Tgfβ2, Stat3 and Thy-1 in native cells is a feature of dedifferentiation in proteomic data (figure legend). **CP** – canonical pathways; Functional annotation $p<0.001$.

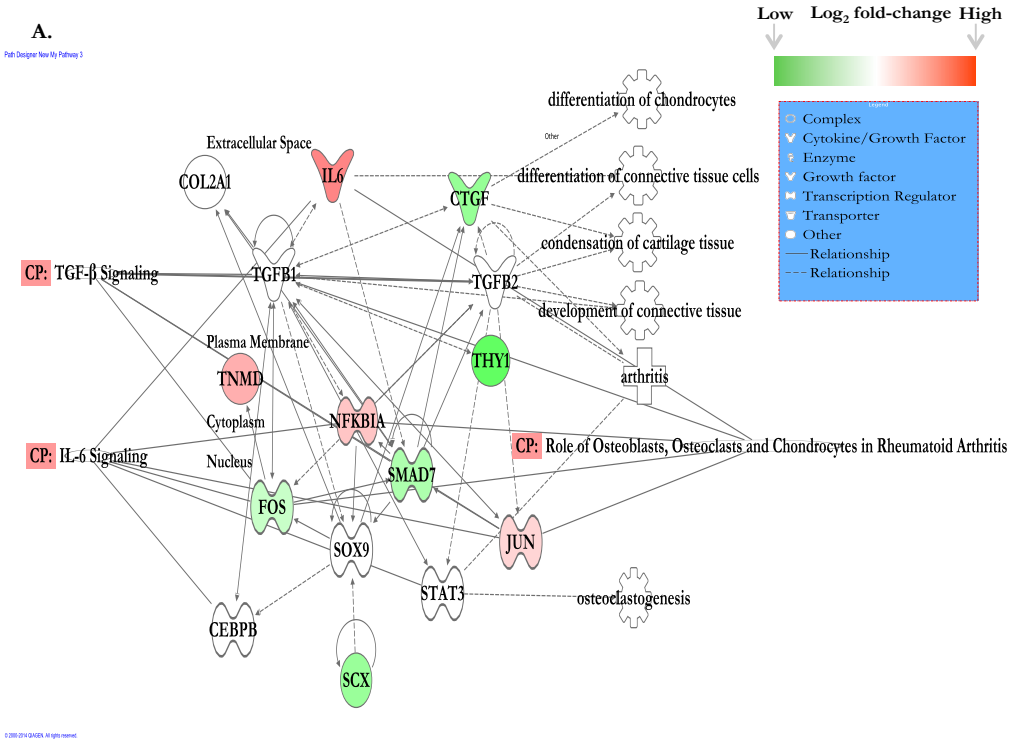


Figure 7.9A: Model network to unify findings from **Figures 7.5-7.7**. Network derived from prediction of Tgf-β1 as a core upstream regulator of expression changes associated with native tenocyte transition to monolayer. Predicted intermediate regulators and tendon markers are overlain with differential expression values from **A**) Affymetrix. Reduction in expression of Ctgf, Scx, Smad7, and Thy-1 in native cells is feature of dedifferentiation in Affymetrix data.

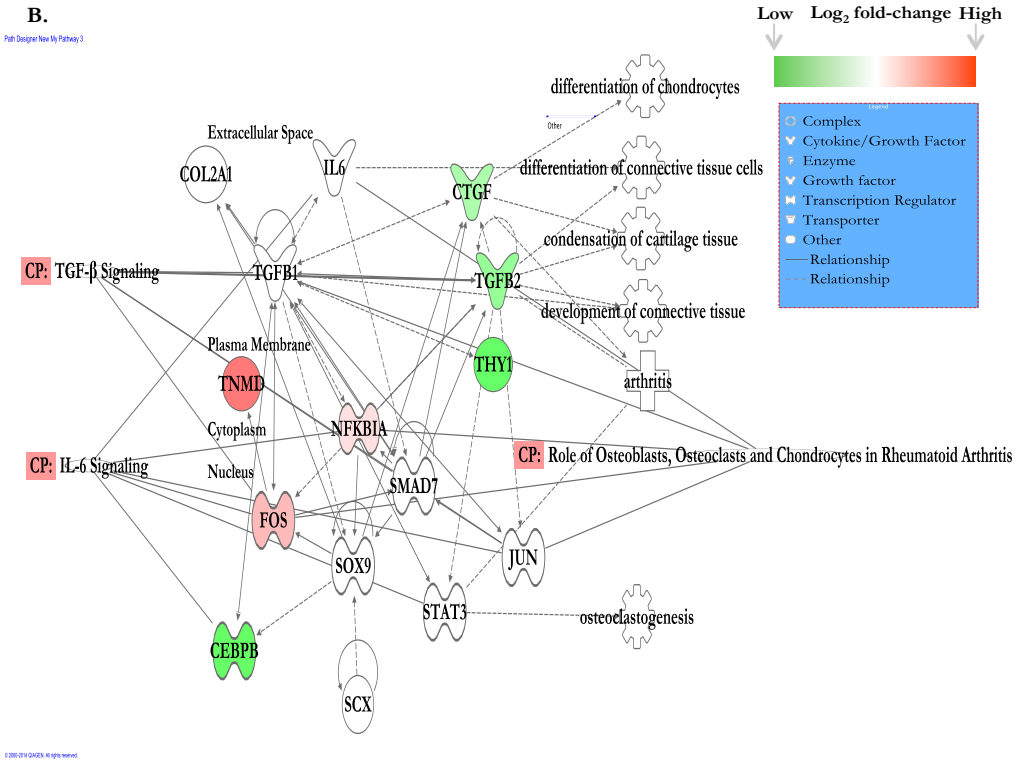


Figure 7.9B: Model network to unify findings from **Figures 7.5-7.7**. Network derived from prediction of Tgf-β1 as a core upstream regulator of expression changes associated with native tenocyte transition to monolayer. Predicted intermediate regulators and tendon markers are overlain with differential expression values from **B**) Illumina. Reduction in expression of Ctgf, Tgf-β2, Cepbp and Thy-1 in native cells is feature of dedifferentiation in this model.

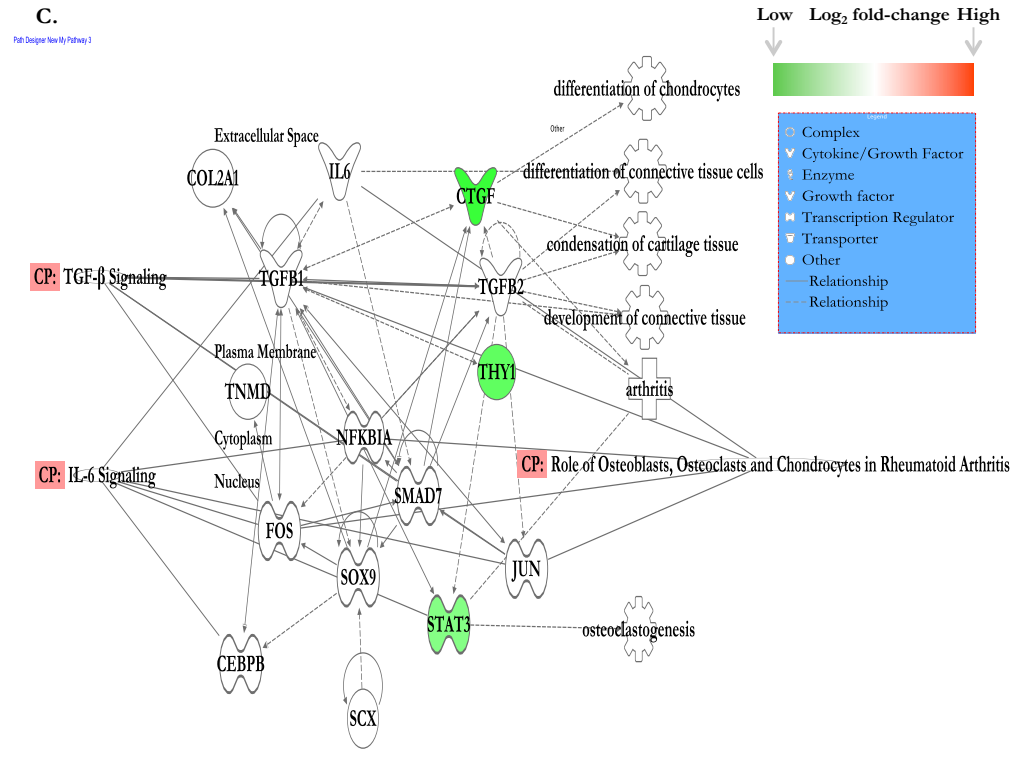


Figure 7.9C: Model network to unify findings from **Figures 7.5-7.7**. Network derived from prediction of Tgf-β1 as a core upstream regulator of expression changes associated with native tenocyte transition to monolayer. Predicted intermediate regulators and tendon markers are overlain with differential abundance values from **C**) proteomic relative quantification study. Reduction in abundance of Ctgf, Stat3 and Thy-1 in native cells is feature of dedifferentiation in proteomic studies – see also **Figure 7.8C**.

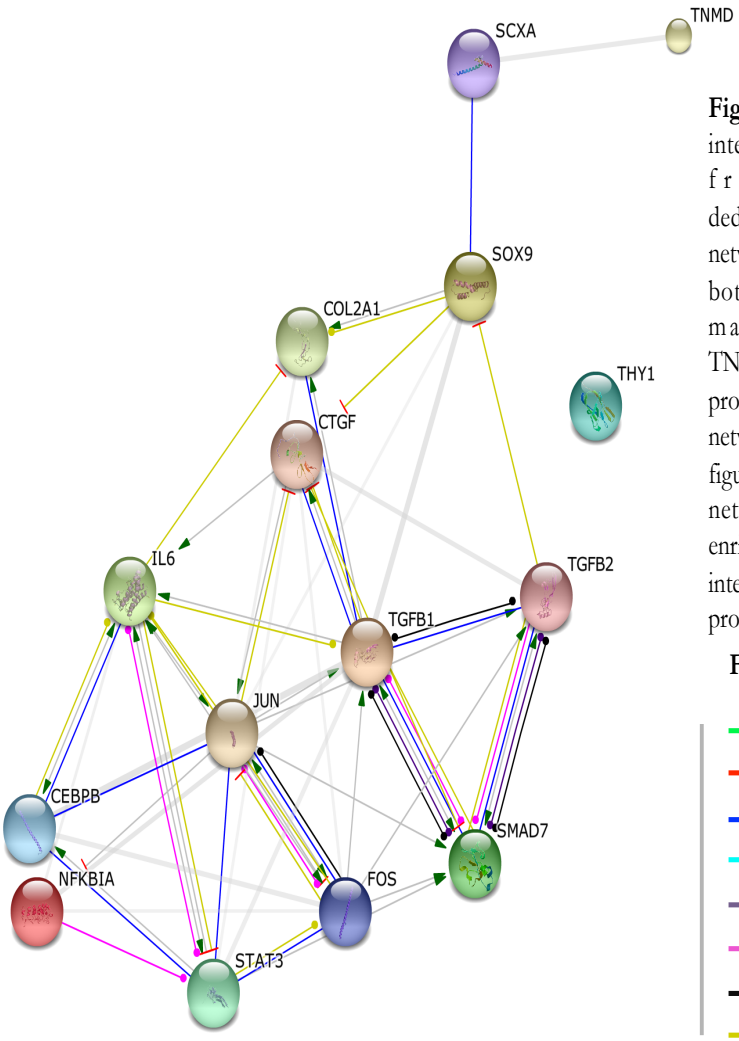


Figure 7.10: Protein-protein interaction network derived from STRING for dedifferentiation mechanistic network. Network considers both cartilage and tendon markers (COL2A1 and TNMD). Modes of actions for proteins are indicated in the network and are defined in the figure legend. Interaction network was found to be enriched (p = 3.4 e-12) with 45 interactions observed on 15 proteins.

Figure legend

- Activation
- Inhibition
- Binding
- Phenotype
- Catalysis
- PTM
- Reaction
- Expression

Table A:

KEGG Pathways	# of Proteins	FDR
Rheumatoid arthritis	5	9.2e-7
TNF signaling pathway	5	2.8e-6
Osteoclast differentiation	5	5.6e-6

Table A: KEGG pathways associated with proposed dedifferentiation network with associated false discovery rate (FDR) and the number (#) of proteins associated with the term

Table B:

Biological Process Terms	# of Proteins	FDR
Regulation of cartilage development	5	3.1e-6
Response to TGFbeta	6	1.2e-5
Cartilage condensation	4	5.6e-6

Table B: Biological process ontology terms associated with proposed dedifferentiation network with associated false discovery rate (FDR) and the number (#) of proteins associated with the term

7.3.4: Re-differentiation transition from monolayer is associated with PDGF BB mediated network

The monolayer to three-dimensional construct transitions were predicted to be associated with the inhibition of PDGF BB, TNF and IL-1B in gene expression studies, and with TGF-β1 activation in proteomics studies for both chondrocytes and tenocytes (**Table 7.2**). Mechanistic networks were prepared for the highest scoring regulators for chondrocytes (PDGF BB) and tenocytes (TNF) and with focused components of the networks shown for IL-6 (chondrocytes) and IL-1B (tenocytes) in **Figures 7.11-7.12**. In proteomics studies JUN was a common intermediate regulator in the TGF-β1 network for both chondrocytes and tenocytes and the down-stream differentially abundant targets are presented in **Figure 7.13**.

Transcriptomic- and proteomic-derived mechanistic networks regulated by PDGF BB and TNF overlapped with a number of intermediate regulators previously described for dedifferentiation including FOS, ApoE, IL-6, JUN, and EGR1.

To define whether these de- and re-differentiation transitions could be unified in a single mechanistic network a model was prepared as before using PDGF BB as the common upstream regulator and including downstream targets present in both de- and re-differentiation networks and differentially expressed genes or proteins frequently identified as targets across the data sets, **Figures 7.14– 7.19**. Protein-protein interactions, **Figure 7.20**, were not significantly enriched, however KEGG pathways were comparable to the dedifferentiation network. Functional annotations were associated with ECM organization and anatomical structure morphogenesis.

SERPINE1 and SERPINE2 show reciprocal expression across data sets

The serine protease inhibitors and chaperones SERPINE1 and SERPINE2 were found to be common downstream targets of predicted regulators across data sets. In the model networks presented in **Figures 7.14** and **7.15-7.19** reciprocal expression patterns were noted with SERPINE2 showing higher gene and protein expression in native and three-dimensional conditions for both chondrocytes and tenocytes, whilst SERPINE1 expression was higher in monolayer for both cell types.

Cell type	Data source	Upstream regulator	Activation *z*-score	Overlap *p*-value	DE genes available
Chondrocytes	Illumina	**PDGF BB**	**-6.13**	1.2e-28	220 (23)
		TNF	-4.85	5.3e-26	276 (16)
		IL1B	-5.88	3.4e-17	222 (12)
Chondrocytes	Affymetrix	TP53	-0.11	1.1e-32	512 (24)
		PDGF BB	**-4.78**	4.6e-30	335 (20)
		IL1B	-6.2	5.3e-22	352 (13)
Chondrocytes	Proteomics	MYC	2.5	3.9e-40	279 (17)
		TP53	1.23	2.5e-33	272 (18)
		TGFB1	**5.5**	5.8e-17	254 (16)
Tenocytes	Illumina	MYC	-0.58	1.6e-13	165 (16)
		TNF	**-2.3**	1.5e-07	211 (16)
		PDGF BB	-2.5	3.9e-07	172 (13)
Tenocytes	Affymetrix	**TNF**	**-6.9**	1.94e-56	266 (15)
		IL1B	-6.84	1.95e-56	250 (11)
		PDGF BB	-6.85	1.04e-48	211 (16)
Tenocytes	Proteomics	MYC	-0.56	5.8e-42	380 (16)
		APP	-2.29	9.1e-30	340 (18)
		TGFB1	**2.6**	7.13e-19	401 (18)

Table 7.2: Top upstream regulators predicted using Ingenuity Pathway Analysis knowledge base from differential expression/abundance for the **monolayer to three-dimensional construct** comparison across three data sources and two cell types. Small molecules and other chemicals predicted by IPA are not included. The activation ***z*-score** indicates the predicted activation status of the mechanistic network. Only genes with available mechanistic networks are shown. **DE genes available** indicates those genes in the differential lists that are known down-stream targets of a number of mechanistic regulators (defined in parentheses). For the transcriptomic data PDGF BB is commonly inhibited; for tenocytes TNF scored higher in both data sets. For the proteomics data results were more variable, however, in both cases TGFB1 mechanistic network was predicted to be activated, the inverse of the native to monolayer comparison.

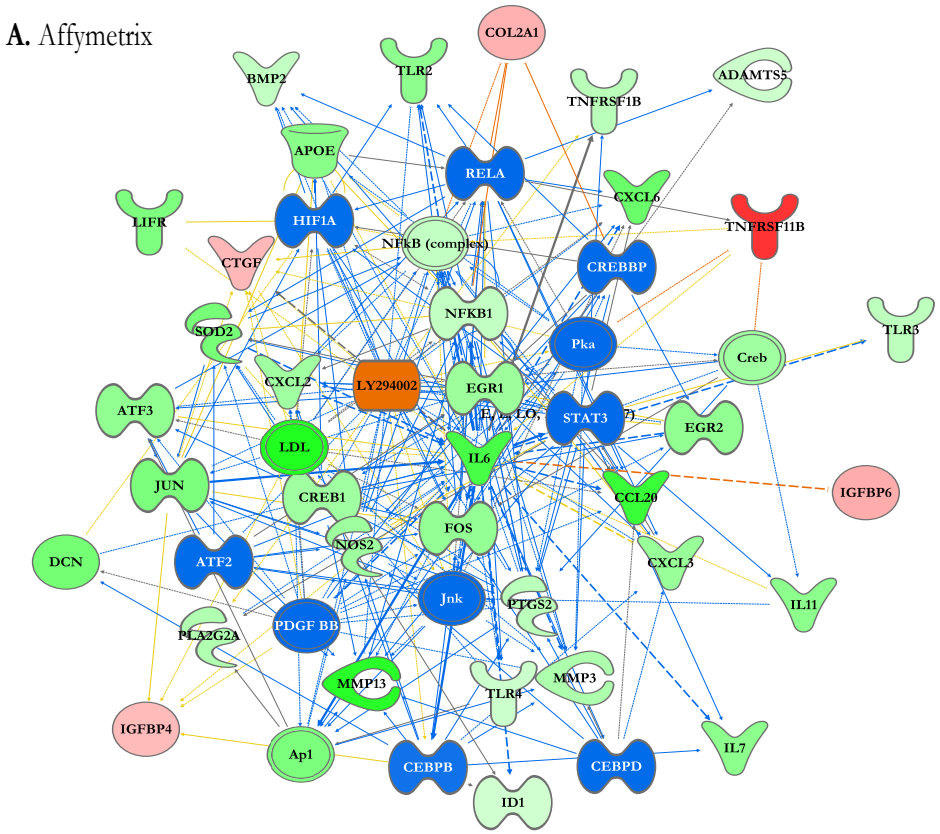
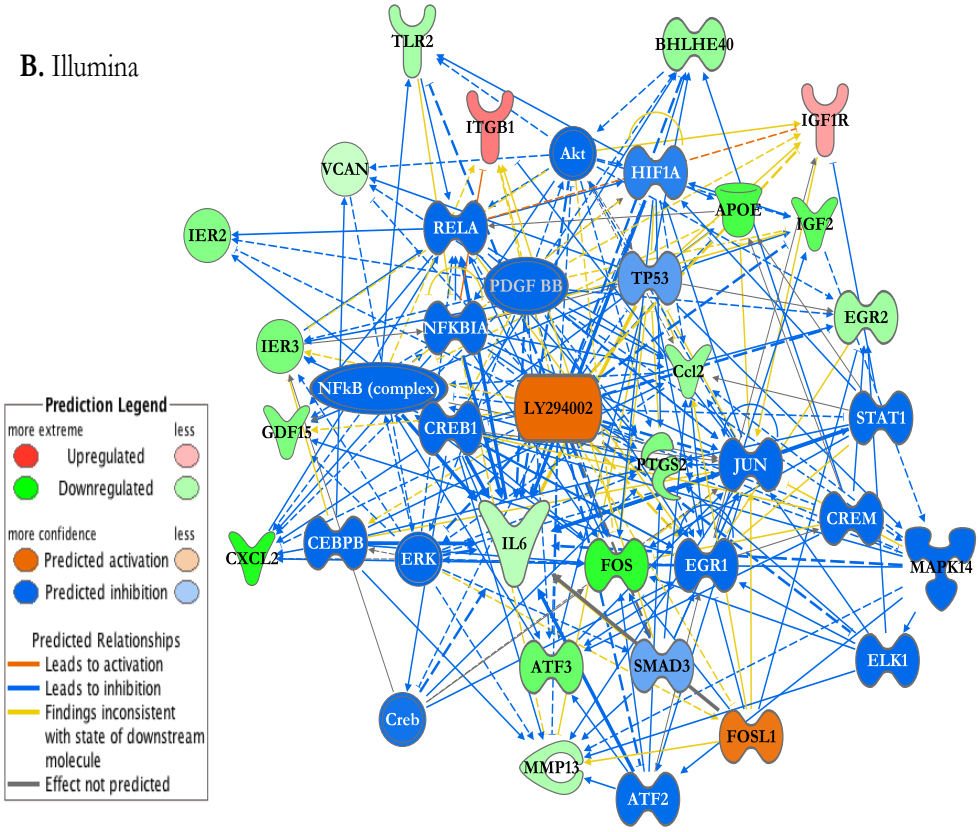


Figure 7.11: Mechanistic network derived from downstream targets of Pdgf bb to show genes with direct interactions with IL-6. Nodes overlaid with differential expression data from **A**: Affymetrix, **B**: Illumina monolayer to alginate comparison. In both gene expression studies lower expression of Il-6, TLR2 and CXCL2 is found in monolayer culture relative to alginate beads under the predicted influence of regulators shown in blue (figure legend). PI3K antagonist LY294002 is shown as activated.



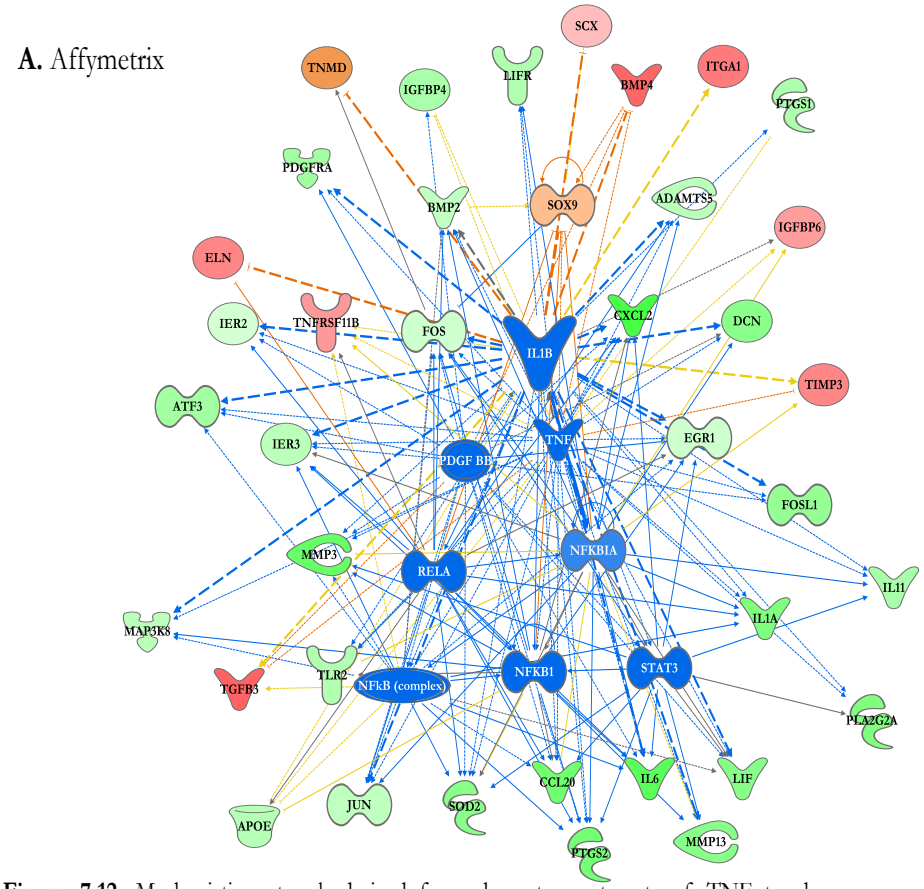
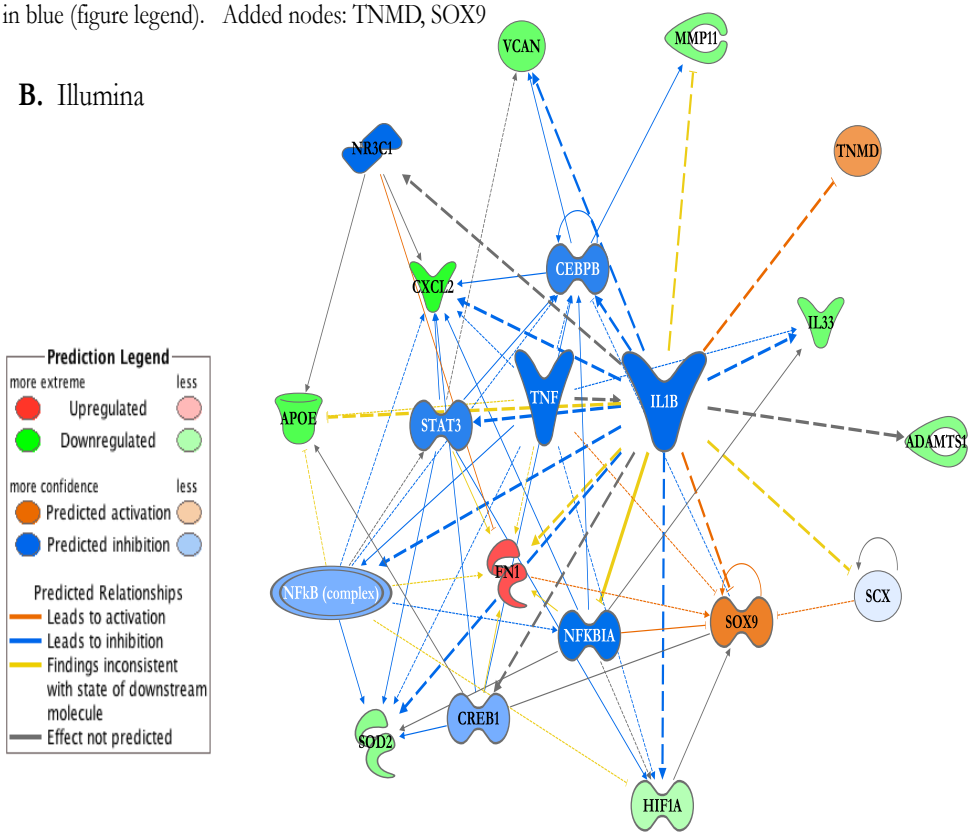


Figure 7.12: Mechanistic network derived from downstream targets of TNF to show genes with direct interactions with IL-1B. Nodes overlaid with differential expression data from **A**: Affymetrix, **B**: Illumina monolayer tenocytes to fibrin comparison. In both gene expression studies lower expression of CXCL2 and ApoE is found in monolayer culture relative to fibrin cultures under the predicted influence of regulators shown in blue (figure legend). Added nodes: TNMD, SOX9



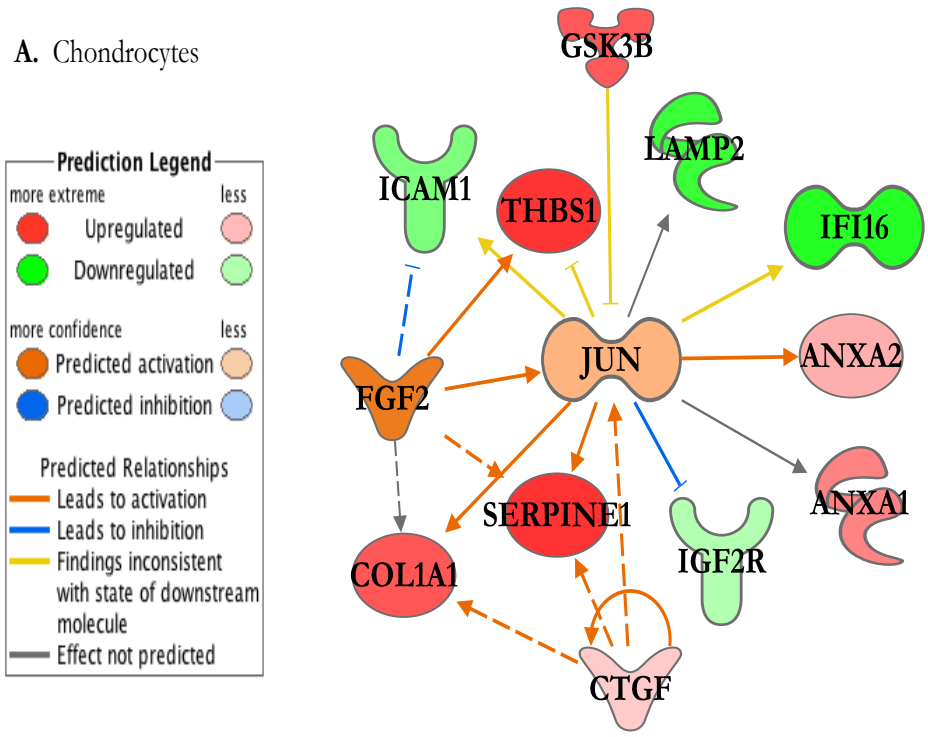
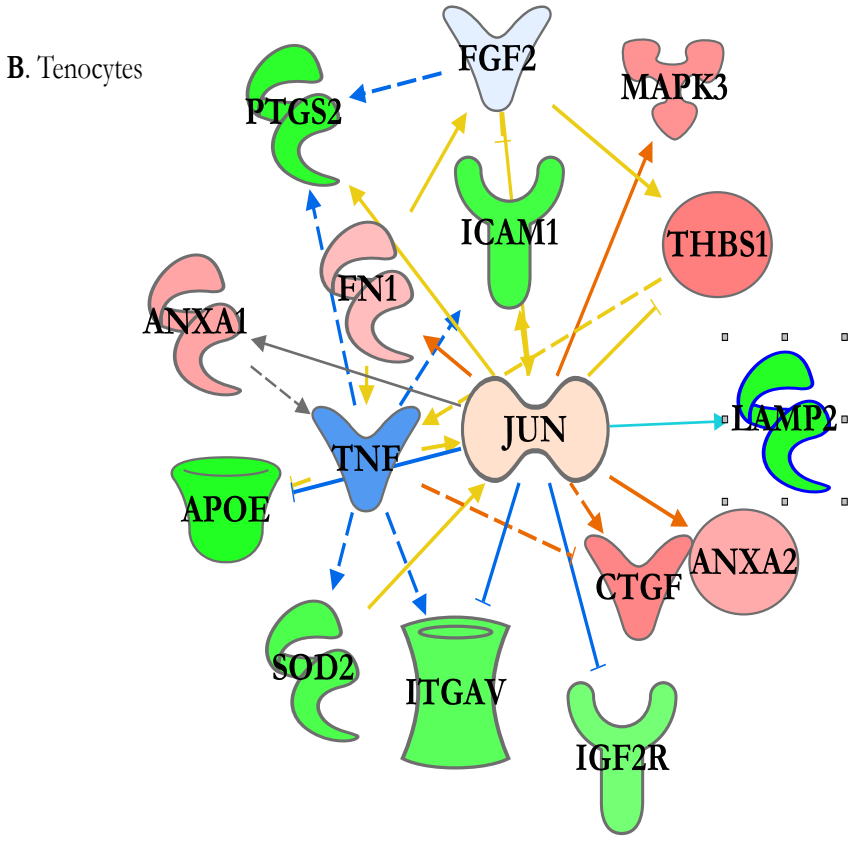


Figure 7.13: Mechanistic network derived from downstream targets of TGF-β1 to show **proteins** with direct interactions with Jun. Nodes overlaid with differential expression data from **A**: monolayer chondrocytes to alginate, **B**: monolayer tenocytes to fibrin cultures. In both protein abundance studies lower levels of ICAM1 and IGF2R are found in monolayer cultures relative to the three-dimensional culture condition (figure legend).



A: Affymetrix: native chondrocytes to monolayer

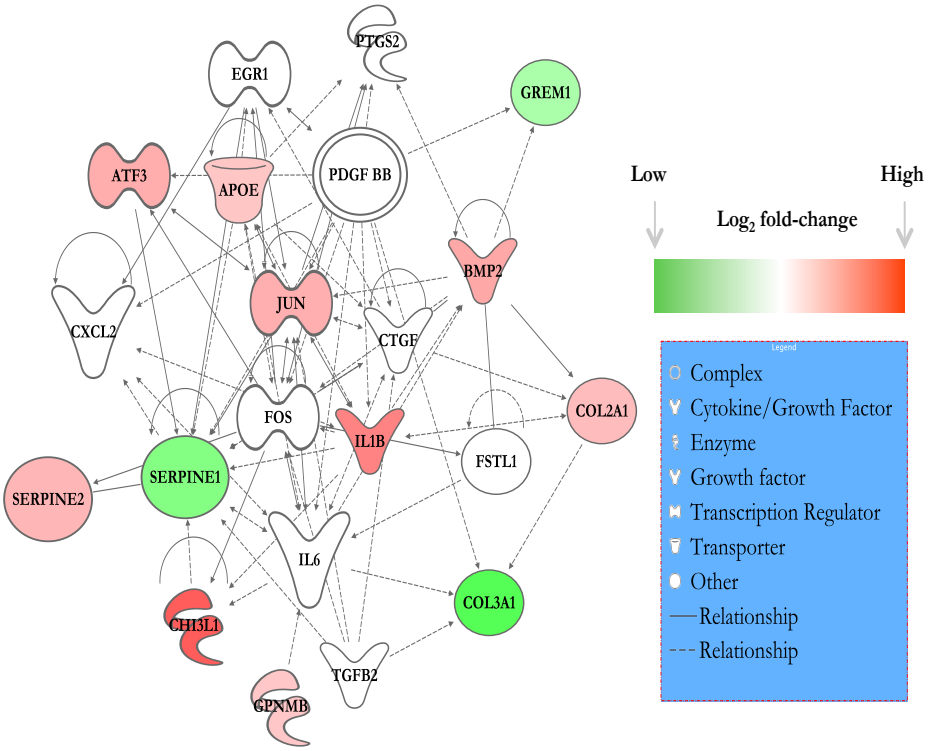
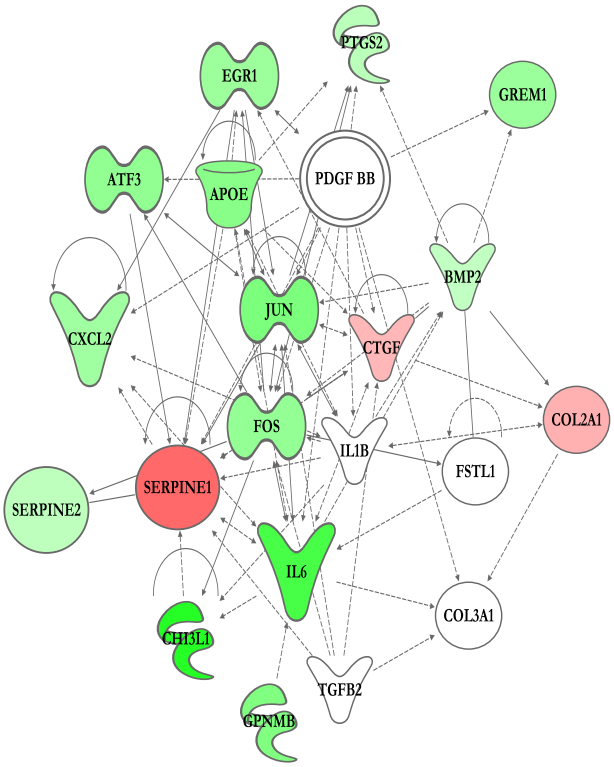


Figure 7.14: Model developed from upstream regulators IL-6, PDGF-bb and IL-1b and genes/proteins with common differential expression across both cartilage and tendon studies. **A**: Higher expression of IL-1b and COL2A1 in native chondrocytes with GREM1, SERPINE1 and COL3A1 more robustly expressed in monolayer cultures (figure legend); **B**: The re-differentiation transition finds a reciprocal relationship in the expression of SERPINE1 and SERPINE2; further alginate cultures show higher expression of GPNMB, CHI3L1 and ApoE relative to monolayer.

B. Affymetrix: monolayer chondrocytes to alginate



A: Illumina: native chondrocytes to monolayer

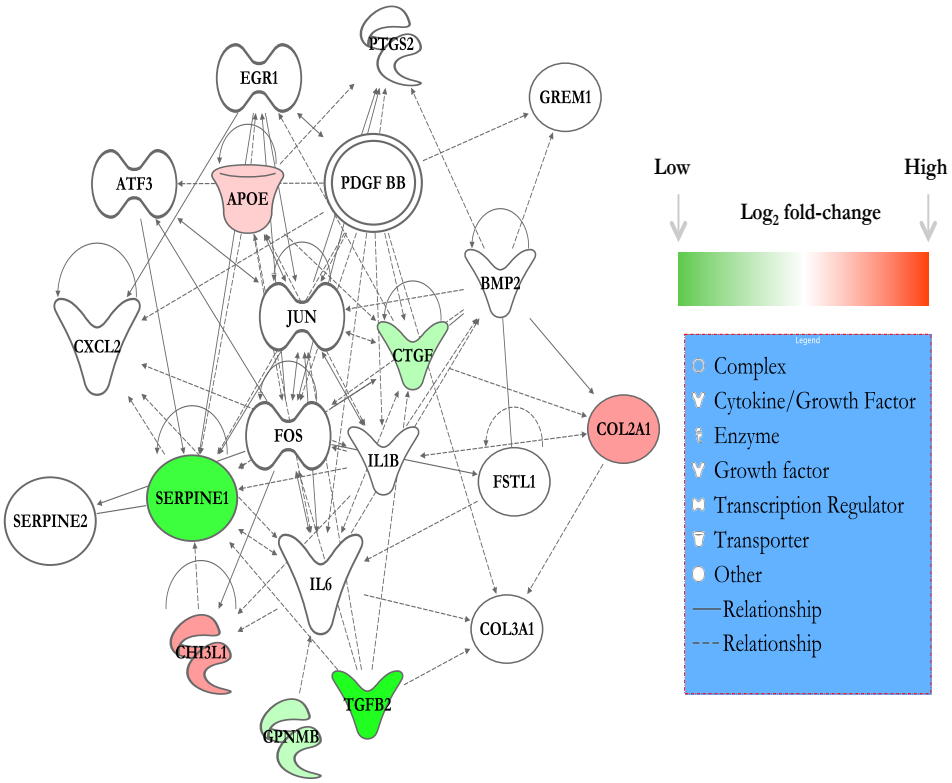
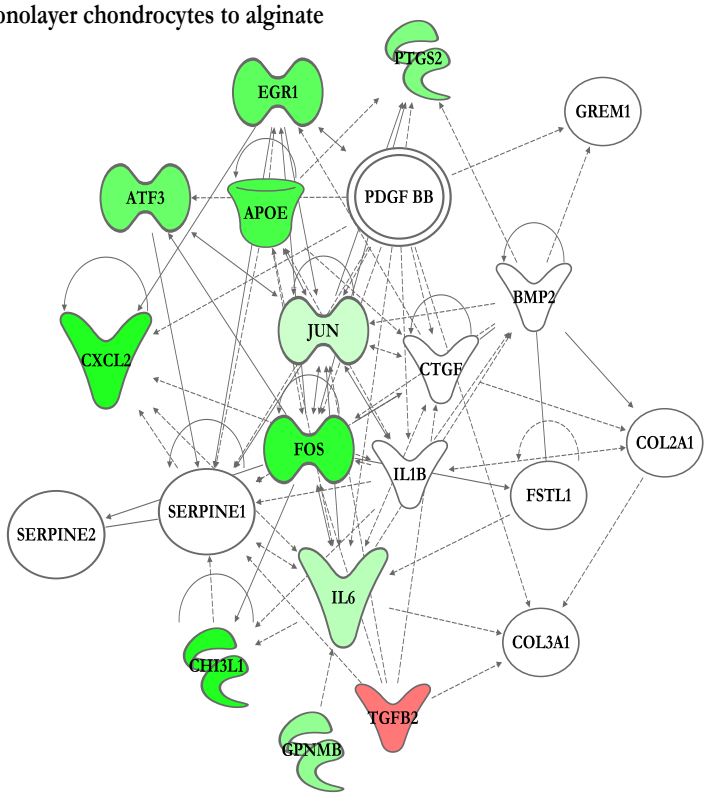


Figure 7.15: Model developed from upstream regulators IL-6, PDGF-bb and IL-1b and genes/proteins with common differential expression across both cartilage and tendon studies. **A**: Higher expression of ApoE and COL2A1 in native chondrocytes with SERPINE1, GPNMB and TGFB2 more robustly expressed in monolayer cultures (figure legend); **B**: The re-differentiation transition finds a pro-inflammatory response with higher expression of IL-6, CXCL2 and PTGS2 in alginate cultures. Monolayer shows lower expression of CHI3L1 and ApoE relative to alginate indicating some restitution of the native expression profile.

B. Illumina: monolayer chondrocytes to alginate



A: Proteomics: native chondrocytes to monolayer

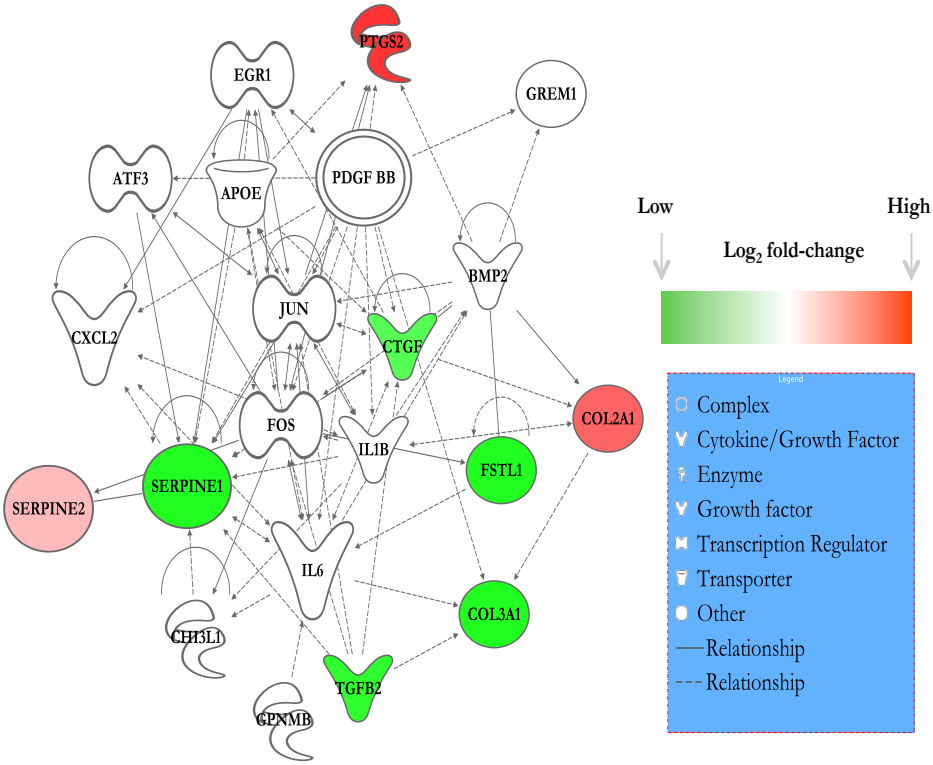
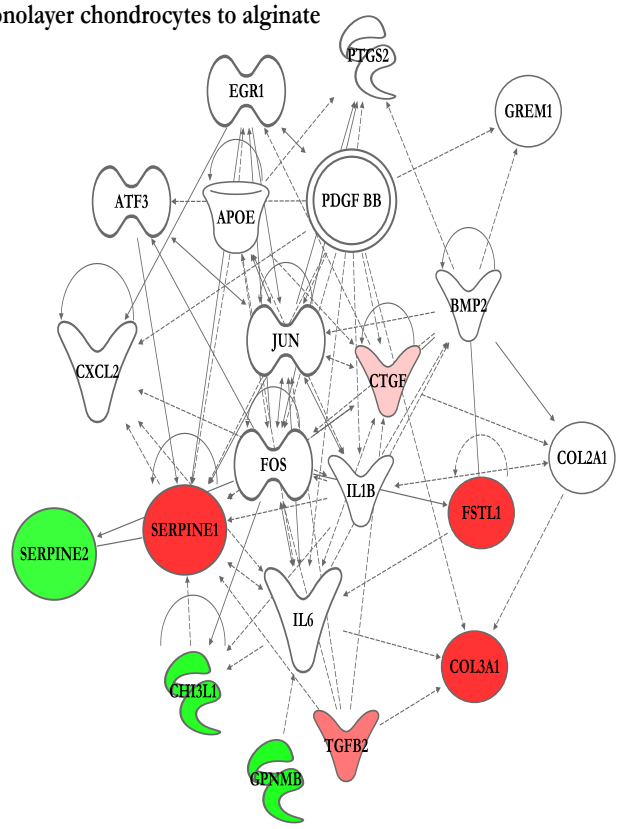


Figure 7.16: Model developed from upstream regulators IL-6, PDGF-bb and IL-1b and genes/proteins with common differential expression across both cartilage and tendon studies. **A**: Higher expression of Col2a1, Serpine2 and Ptgs2 in native chondrocytes with Serpine1, Ctgf and Tgfb2 more robustly expressed in monolayer cultures (figure legend); **B**: The re-differentiation transition finds the reciprocal relationship between Serpine1 and Serpine2 demonstrable in the protein profile of chondrocytes.

B. Proteomics: monolayer chondrocytes to alginate



A: Affymetrix: native tenocytes to monolayer

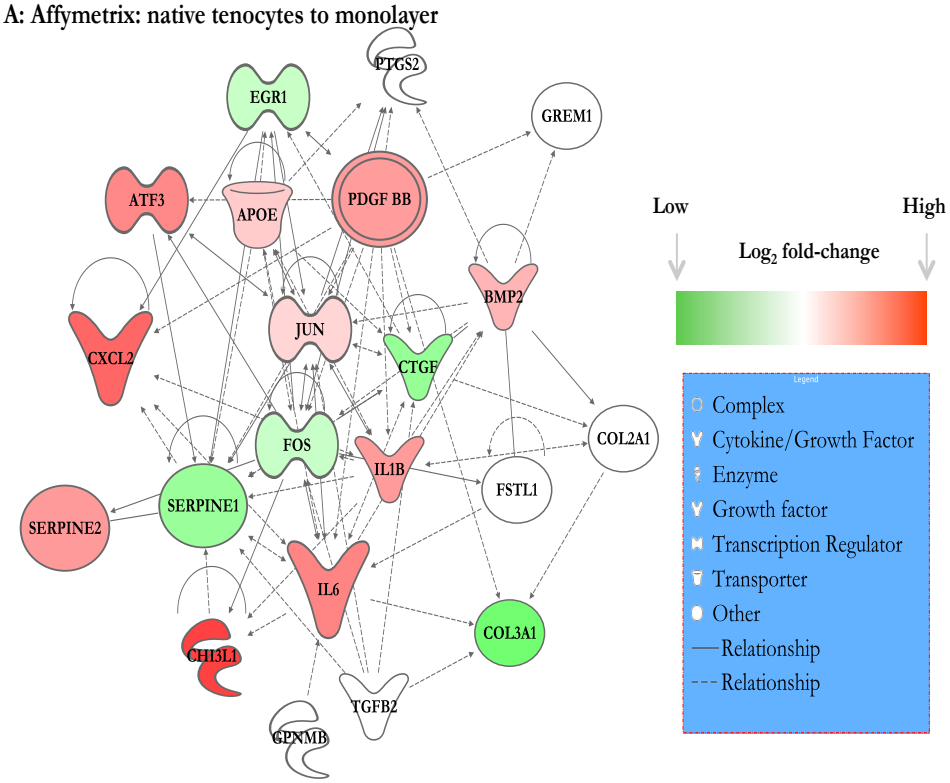
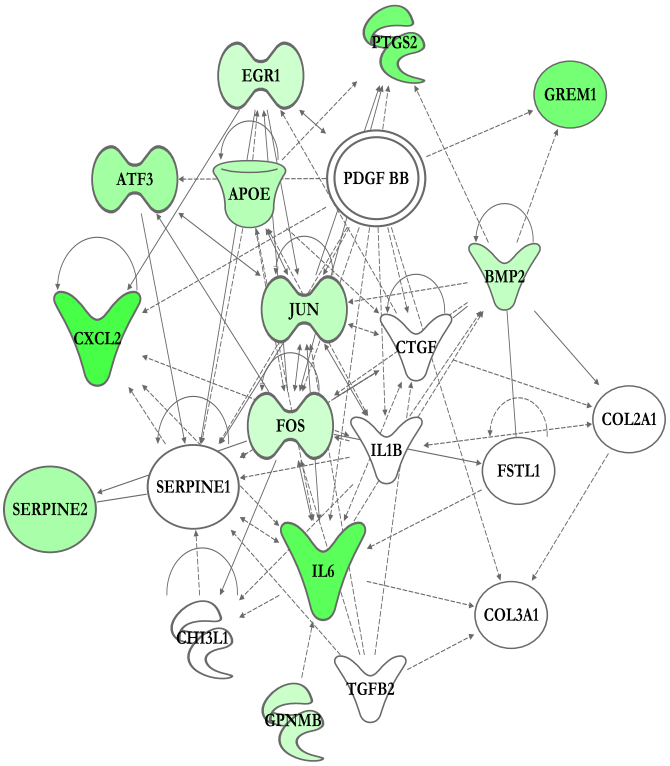


Figure 7.17: Model developed from upstream regulators IL-6, PDGF-BB and IL-1b and genes/proteins with common differential expression across both cartilage and tendon studies. **A**: Higher expression of key regulators PDGF BB, IL-6 and IL-1b in native tenocytes (figure legend); **B**: The re-differentiation transition finds lower expression of IL-6, SERPINE2, BMP2 and ATF3 in monolayer relative to fibrin constructs suggestive of an expression profile consistent with some restitution of a differentiated state.

B. Affymetrix: monolayer tenocytes to fibrin constructs



A: Illumina: native tenocytes to monolayer

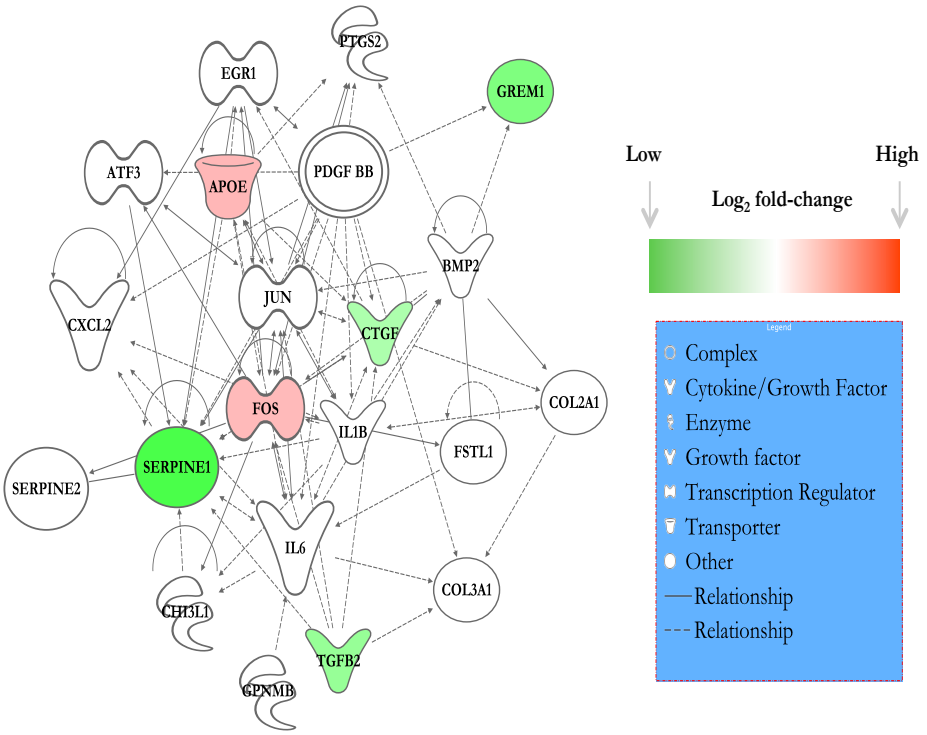
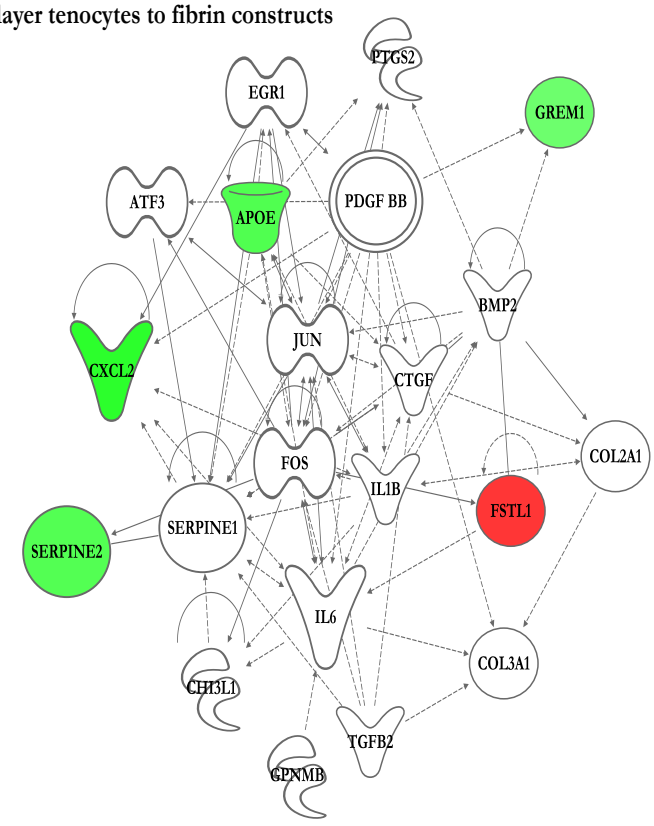


Figure 7.18: Model developed from upstream regulators IL-6, PDGF-BB and IL-1b and genes/proteins with common differential expression across both cartilage and tendon studies. **A**: Higher expression of ApoE and FOS in native tenocytes is associated with reduced expression of CTGF and GREM1 relative to monolayer (figure legend); **B**: The re-differentiation transition finds lower expression SERPINE2 and GREM1 in monolayer relative to alginate.

B. Illumina: monolayer tenocytes to fibrin constructs



A: Proteomics: native tenocytes to monolayer

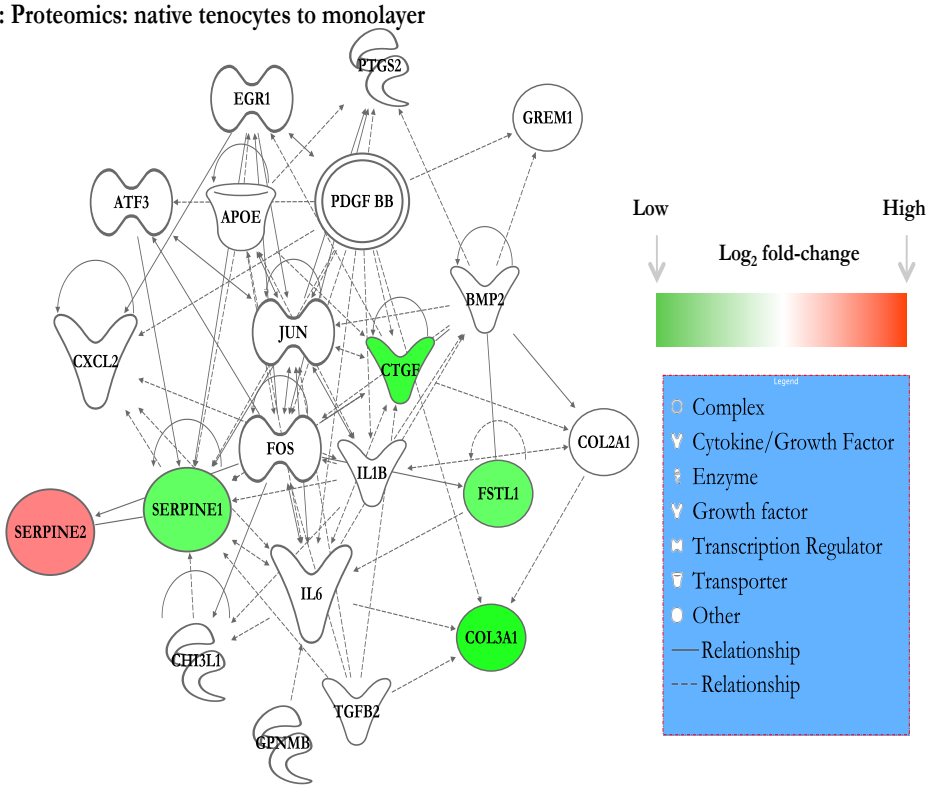
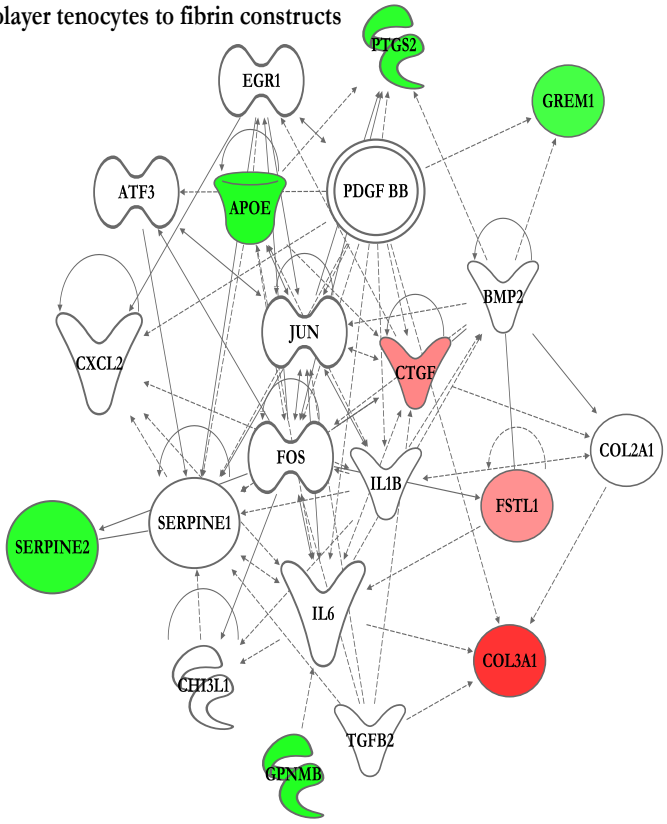


Figure 7.19: Model developed from upstream regulators IL-6, PDGF-bb and IL-1b and genes/proteins with common differential expression across both cartilage and tendon studies. **A**: Reciprocal expression of SERPIN proteins was evident in dedifferentiation with higher abundance of SERPINE2 in native tenocytes (figure legend); **B**: The higher expression of FSTL1, CTGF and COL3A1 was evident in monolayer with lower expression of SERPINE2 relative to fibrin.

B. Proteomics: monolayer tenocytes to fibrin constructs



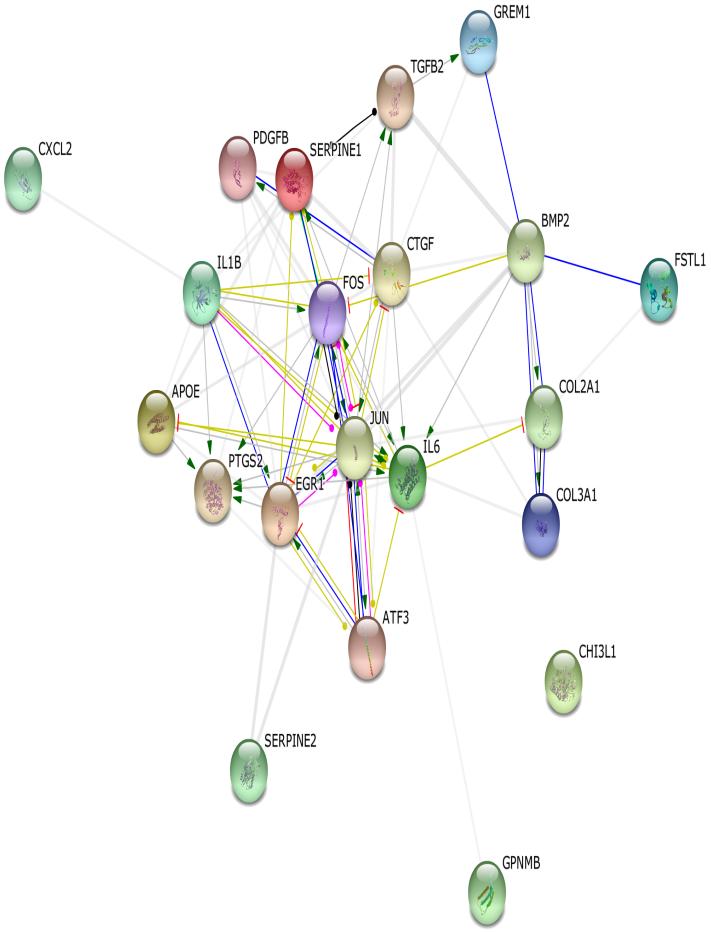


Figure 7.20: Proposed universal mechanistic network for de- and re-differentiation. Protein-protein interaction network derived from STRING. Modes of actions for proteins are indicated in the network and are defined in the figure legend. Interaction network was not found to be enriched with 74 interactions observed on 21 proteins.

Table A:

KEGG Pathways	# of Proteins	FDR
TNF signaling pathway	6	2.5e-7
Rheumatoid arthritis	5	6.1e-6
Osteoclast differentiation	4	1.9e-3

Table A: KEGG pathways associated with proposed dedifferentiation network with associated false discovery rate (FDR) and the number (#) of proteins associated with the term

Table B:

Biological Process Terms	# of Proteins	FDR
Response to mechanical stimulus	9	1.1e-9
Extracellular matrix organisation	8	8.04e-7
Anatomical structure morphogenesis	12	1.2e-5

Table B: Biological process ontology terms associated with proposed dedifferentiation network with associated false discovery rate (FDR) and the number (#) of proteins associated with the term

7.4: Discussion

7.4.1: Union of data sets and functional annotations

There was variable overlap between data sets in terms of commonly differentially represented genes and proteins. Only a fraction of genes expressed in either the Illumina or Affymetrix studies were represented in the proteomics studies. Additionally, across all studies there were consistently very few common elements. For transcriptomic data this not unexpected (Manoli, Gretz et al. 2006). Using only these common elements to integrate functional annotations would encourage erroneous and biased descriptors of the data. The integrated annotations presented here represent a consensus across two data sets using the MONA algorithm (Sass, Buettner et al. 2013, Sass, Buettner et al. 2014), which uses a Bayesian approach to determine marginal posteriors for the gene ontology terms. This allowed the simultaneous assessment of data sets arising from both mRNA and protein studies.

Integrated biological process functional annotations across transcriptomic and proteomic data sets demonstrated general concordance with findings previously presented. In particular the relevance of metabolic changes associated with oxidation-reduction, carbohydrate and phospholipid biosynthetic processes were prominent. A number of development-associated terms relating to muscle structure, vasculature, and cardiovascular development and cell morphogenesis were found across data sets. Muscle-associated functional annotations were not restricted to tendon-derived data sets. The application of this omics integration technique also highlighted Wnt- and NF-κB signalling, which although evident in the differential expression analysis in individual data sets, have not been consistently identified as significant gene ontology annotations.

The actions of Il-6-mediated signaling have been discussed in **Chapter 5**. Discussion here focuses on TGF-β-mediated mechanisms related to dedifferentiation.

7.4.2: Shared upstream regulators guide mechanistic networks

Union of elements or ontology-based approaches do not provide a mechanistic insight into the system under investigation; neither do they directly make use of the quantitative data. In order to explore whether common mechanisms were regulating the dedifferentiation or redifferentiation phenotypes Ingenuity Pathway Analysis was used to define upstream regulators shared between chondrocytes and tenocytes. Using known protein-protein interaction databases it was possible to demonstrate a highly enriched TGF-β centric network associated with dedifferentiation. Further Ingenuity analysis of the redifferentiation transition predicted a reciprocal relationship between TGF-beta activation in monolayer and PDGF activation in three-dimensional cultures.

SMAD7 as a mediator of differentiation status

In musculoskeletal tissues TGF-βs are diverse regulators of differentiation steering developmental programs such as cell-fate decisions between cartilage and tendon (Lorda-Diez, Montero et al. 2009), chondrogenic condensation and proliferation (Finnson, Chi et al. 2012) and are key regulators of disease. In a mechanical disruption osteoarthritis model high concentrations of TGF-β1 were found in subchondral bone (Zhen, Wen et al. 2013). In this study transgenic activation of TGF-β1 expression induced osteoarthritis, whilst knock-out of the TGF–β type II receptor in subchondral MSCs mitigated these effects. As such the TGF-β-family

includes rational targets for regenerative interventions e.g. gene-based therapy (Madry and Cucchiarini 2013, Zhen and Cao 2014).

That TGF-β-mediated signalling is predicted as a master upstream regulator of the gene and protein responses seen in de- and re-differentiation is unsurprising given the considerable complexity of ligand-receptor interactions and plasticity of TGF-β-signalling processes (Schmierer and Hill 2007, Cellière, Fengos et al. 2011). These data sets provide no indication of the temporal nature of the signal (whether transient, sustained or oscillatory) that would define more accurately the differences in cellular response observed across the different culture conditions. Furthermore, a number of contextual determinants, for example factors regulating signal transduction (ligand isoforms), transcription factors that bind to SMAD proteins, and the epigenetic status of the cell, all modulate the TGF-β signaling effects (Massagué 2012). Consequently, analysis of isolated and static gene expression profiles alone cannot provide a consistent understanding of the underlying mechanism when trying to establish the differentiation status of a cell. In brief, the binary nature of pathway predictions in this study (activated *versus* inactivated) resulted in conflicted reports of pathway involvement. This calls into question the reproducibility of pathway predictions and the relevance of predicted intermediate regulators.

Although conflict exists there is still likely to be strong concordance in the biological process involved. With this in mind integration in this chapter considered only common regulators identified across multiple data sets and focused only on downstream effectors that appeared in two or more data sets from the same cell type. Whilst a cautious approach it is qualified by evidence in the literature

indicating the poor reproducibility of 'biosignature' approaches to defining phenotypes arising from small data sets (Azuaje, Zheng et al. 2011).

This analysis sought to focus on the TGF-β-antagonist SMAD7 for which there was evidence of differential expression in dedifferentiation in both chondrocyte and tenocyte analysis across two platforms. Additionally, it was associated with a number of common downstream targets, including CCN2/CTGF and SERPINE1.

SMAD7 can inhibit both BMP (R-SMADS 1/5/8) and TGF-β (R-SMAD 2/3) signaling by competitively binding receptor R-SMADS (Nakao, Afrakhte et al. 1997, Massagué, Seoane et al. 2005), and through ubiquitin-mediated degradation of activated TGF-β receptor complex in association with SMURF2 (Kavsak, Rasmussen et al. 2000). It is implicated in multiple physiological and pathophysiological contexts including embryonic development, fibrosis, tumourigenesis and inflammation (Zhu, Chen et al. 2011) consistent with TGF-β-signaling pathways. Two adenoviral over-expression studies of SMAD7 (from the same group) have been shown to be chondroprotective in osteoarthritis models (Scharstuhl, Vitters et al. 2003, Blaney Davidson, Vitters et al. 2006).

SMAD7 influences differentiation and proliferation in chondrocytes

Effects of SMAD7 on chondrocyte phenotype have been previously established. SMAD7 is expressed in developing growth plates (Sakou, Onishi et al. 1999). In SMAD7 knock-out mice defects in terminal chondrocyte maturation are found with reduced shortened hypertrophic zones in a study of endochondral ossification (Estrada, Wang et al. 2013). Elevated levels of HIF1α were also found in the proliferative zones of SMAD7$^{-/-}$ mice suggesting an impact on the hypoxic stress response in developing cartilage. Further evidence for the effects of SMAD7 on the

differentiation status of chondrocytes comes from adenoviral over-expression of SMAD7 severely inhibited Mekel's cartilage development in mice (Ito, Bringas et al. 2002) whilst SMAD7-associated inhibition of differentiation and proliferation of osteoblasts has also been reported (Yano, Inoue et al. 2012).

Recent work in β-islet cell differentiation in the pancreas (see **Chapter 1**) has demonstrated that reversal of dedifferentiation may be achieved through TGF-β inhibition; Blum, *et al* (2014) presented evidence indicating that an ALK5/TGF-βR1 inhibitor was able to rescue dedifferentiated β-cells and inhibit cytokine-induced β-cell stress (Blum, Roose et al. 2014). Furthermore, enhanced expression of SMAD7 promotes β-cell proliferation *in vivo* (Xiao, Gaffar et al. 2014), and this proliferation in response to β-cell loss may require an initial transition through a dedifferentiation state (El-Gohary, Tulachan et al. 2014).

In summary, SMAD7 modulation of TGF-β signalling in dedifferentiation of chondroyctes, and tenocytes, is predicted in bioinformatics analysis and demonstrated in differential expression studies. TGF-β signalling is implicated both in cartilage disease and development. Evidence indicates that the developmental paradigm is recapitulated in osteoarthritis (Tchetina 2011). It is, therefore, exigent that a better understanding of the dedifferentiation mechanism it developed. In particular defining whether the resulting loss of functional phenotype is firstly, associated with cartilage degeneration, secondly, a regenerative mechanism that may be harnessed.

7.4.3: SERPINE1, an ALK5 target, shows reciprocal expression with SERPINE2 between native, monolayer and three-dimensional cultures

The model networks for de- and re-differentiation data from mRNA and protein studies show the reciprocal expression of two SERPIN family E-clade members, SERPINE1/PAI-1 (plasminogen activator inhibitor 1) and SERPINE2/PN-1 (glial-derived nexin). The latter shows higher expression in the native and three-dimensional conditions, whilst the former shows up-regulation in monolayer. Although previously defined exclusively as serine protease inhibitors most SERPIN-family members are often chaperones with diverse roles (Heit, Jackson et al. 2013).

TGF-β signaling induces SERPINE1 expression through the co-operative binding of R-SMADs and AP-1 (Fos/Jun heterodimer) to the promoter (Zhang, Feng et al. 1998, Guo, Inoki et al. 2005). The concomitant protein PAI-1, a secreted glycoprotein, is the inhibitor of urokinase plasminogen activator (uPA) and tissue plasminogen (tPA), which are required for the conversion of plasminogen to plasmin, a potent proteolytic enzyme (Małgorzewicz, Skrzypczak-Jankun et al. 2013) that activates a number of MMP family members with roles in ECM remodeling. As such PAI-1 titrates ECM degradation permitting the accumulation of ECM components at injury sites. Although primarily an inhibitor of fibrinolysis it is also associated with a number of pathological processes especially fibro-proliferative disorders, renal fibrosis, atherosclerosis, vascular thrombosis and rheumatoid arthritis under the influence of TGFβ1; these activated genes SERPINE1 and CCN2 (see below) are also ROS-dependent (Samarakoon, Overstreet et al. 2013).

SERPINE2/PN-1 functions are less well understood, however, it has been shown to inhibit a number of serine proteases including thrombin and urokinase;

expression in astrocytes, endothelial cells and fibroblasts is reported and is associated with the ECM. It is also defined as a substrate for MMP9. It is the only SERPIN found at physiological levels in the brain (Sappino, Madani et al. 1993). Recent evidence associated SERPINE2 expression with pro-neoplastic properties (Bergeron, Lemieux et al. 2010, Wang, Wang et al. 2014) including increased ECM deposition (Buchholz, Biebl et al. 2003). SERPINE2 risk alleles associated with chronic obstructive pulmonary disorder (COPD) are widely reported (Demeo, Mariani et al. 2006, Zhu, Warren et al. 2007). Additionally, SERPINE2 has also been shown to have antagonistic effects on Hedgehog signaling (Vaillant, Michos et al. 2007, McKee, Xu et al. 2012). Whether modulation of SERPIN E protease inhibitors is relevant to the differentiation status of model cultures has not been determined, however it may be related to modulation of the forming ECM in these model cultures.

7.4.4: CCN2/CTGF, a differentiation regulator and pro-fibrosis mediator, is associated with dedifferentiation

Evidence is presented for the higher expression of the CCN family member CCN2/CTGF, connective tissues growth factor, in dedifferentiated chondrocytes and tenocytes at the gene and protein level. In **Chapter 1** the relevance of CCN-family members to the ECM is highlighted.

CCN family members are immediate-early growth-responsive genes (with the exception of CCN3/NOV) with roles in cell proliferation, differentiation, embryogenesis and wound healing. Dysregulated expression is extensively associated with a number of fibro-degenerative and fibro-proliferative conditions including ECM deposition in atherosclerosis, lung and kidney fibrosis, scleroderma and muscular dystrophies (Oemar and Lüscher 1997, Leask, Parapuram et al. 2009,

Morales, Gutierrez et al. 2013). CCN2 also shows osteoarthritis-associated pro-inflammatory effects (Wang, Qiu et al. 2013); intense CCN2 staining by IHC and ISH is evident in proliferative chondrocytes in moderate to severe osteoarthritis (Omoto, Nishida et al. 2004). The presence of CCN2 proteolytic fragments in the media of femoral head explants of normal cartilage is reported (Wilson, Whitelock et al. 2009) and this increases in the presence of IL-1 (Wilson, Belluoccio et al. 2008).

CCN2 is a cysteine-rich secreted protein with a high amino-acid homology with other CCN family members. CCN proteins have a modular structure consisting of four-conserved modules: insulin-like growth factor binding protein-like (IGFBP) module, b) Williebrand factor type C (VWC) module; c) thrombospondin (TSP) type I repeat, and d) C-terminal module with ECM protein, growth factor and integrin binding potentials (Takigawa 2013). The C-terminal module is homologous to *slit* (Oemar and Lüscher 1997) (see **Chapter 1 and 2**) associated with axonal guidance and tendon development in *Drosophila* (Schweitzer, Zelzer et al. 2010). Notably, CCN2 mRNA is induced by TGF-β, but not by PDGF, EDGF, or bFGF (Igarashi, Okochi et al. 1993); a TGF-β responsive consensus sequence has been identified in the CCN2 promoter region (Grotendorst, Okochi et al. 1996).

CCN2 in endochondral ossification

Early research considered CCN2 exclusively in the context of TGF-β1 mediated fibrotic conditions; further work demonstrated the pro-hypertrophic effects of CCN2 on growth-plate chondrocytes (Nakanishi, Nishida et al. 2000). Comparable promotion of differentiated status was subsequently found on osteoclasts and vascular endothelial cells defining CCN2 as 'ecogenin' or endochondral ossification genetic factor for its manifold effects on the proliferation and differentiation of

growth-plate chondrocytes, osteoblasts and endothelial cells (Ivkovic, Yoon et al. 2003, Takigawa 2013).

Transgenic mice generated to over-express CCN2/*lacZ* fusion gene under the control of the COL2A1 promoter were found to be, relative to the wild-type, larger, with longer tibia, there was evidence of enhance chondrocyte proliferation and increased proteoglycan and collagen type II accumulation in the proliferative and resting zones (Tomita, Hattori et al. 2013).

CCN2 in differentiation status

There is normally restricted expression of CCN2 in the adult (usually hepatic stellate cells and kidney mesangial cells) and the induction of CCN2 expression by TGF-β is generally restricted to cells of mesenchymal origin (Leask, Parapuram et al. 2009). CCN2 may be self-regulating in mesenchymal cells by the sustaining particular actions of TGF-β (Grotendorst 1997) probably through mediating levels of SMAD7 (Qi, Chen et al. 2007, Sobral, Montan et al. 2011, van Rooyen, Schäfer et al. 2013). These reports are consistent with findings in this study of chondrocytes and tenocytes in monolayer demonstrating a TGFβ-CCN2 mediated regulatory network with elevated SMAD7 likely to represent a modulatory influence.

It is not clear how the higher expression of CCN2 should be interpreted in the context of dedifferentiation as evidence relates both to musculoskeletal developmental stages and fibroblast proliferation mechanisms. In gain of function studies investigating signaling mechanisms regulating cartilage and tendon differentiation, local application of TGF-β loaded micro-beads to the undifferentiated inter-digital mesenchyme in avian limb buds promotes the expression of CCN2, followed by SOX9 and BMPR1b genes; this is contrasted in

the digit tip, where chondrogenesis is inhibited and TGF-β application results in over-expression of scleraxis (SSCX) driving a tendon phenotype (Lorda-Diez, Montero et al. 2014).

CCN2 has also been demonstrated to differentiate mesenchymal stem cells into fibroblasts (Lee, Shah et al. 2010) with enhanced expression of collagen type I and tenascin-C; treated cells had an attenuated capacity for tri-lineage differentiation potential into osteogenic, chondrogenic or adipogenic cells.

The reported evidence gives conflicting accounts of the differentiation effects of CCN2 and these effects may be cell-restricted. A definitive understanding of the actions of CCN2 on the differentiation status of chondrocytes and tenocytes in monolayer culture is not possible from these data sets alone.

CCN2 in energy production

Proliferation and differentiation may appear as mutually exclusive mechanisms but CCN2 promotes both in chondrocyte development. The findings presented in this chapter, of enhanced CCN2 mRNA and protein expression in monolayer and subsequent reductions in re-differentiation, are not at odds with the apparently "contradictory" (Kubota and Takigawa 2015) actions of CCN2 on chondrocyte proliferation and dedifferentiation. In work considering the regulatory effects of CCN2 on energy metabolism in chondrocytes Maeda-Uematsu, and co-workers, (2014) found that CCN2-null mice had an generalised reduction in the level of glycolytic metabolites and intracellular ATP relative to wild-type (WT) controls (Maeda-Uematsu, Kubota et al. 2014). Recovery of intracellular ATP levels in chondrocytes from CCN2-null cells was achieved through the addition of exogenous CCN2. Conversely ATP levels were reduced in wild-type cells using an

siRNA directed at CCN2 transcripts. The authors concluded that CCN2-mediated ATP regulation was likely to be mediated via the anaerobic pathway. These findings may explain the alterations in CCN2 expression in monolayer and model culture systems parallel to metabolic changes; elevation of CCN2 expression in rapidly proliferating chondrocytes and tenocytes in monolayer culture may be associated with the increased energy demands consistent with findings in the work of Maeda-Uematsu, *et al* (2014).

CCN2 and integrin-signaling

A consideration of CCN2 and integrin-mediated signaling concludes this discussion. As define in **Chapter 1** the CCN-family have core roles in the transduction of extra-cellular signals into the cell *via* integrin-mediated signaling (Leask and Abraham 2006). The roles of integrins in cartilage homeostasis and disease have also been described (Loeser 2014). The C-terminal domain of CCN2 is known to interact directly with fibronectin through integrin alpha5-beta1 (Hoshijima, Hattori et al. 2006). Concentration-dependent increases in the expression of IL-6 in chondrocytes induced by CCN2 is attenuated by an alpha5-beta1 integrin neutralising antibody in a study of synovial fibroblasts from osteoarthritic joints (Liu, Hsu et al. 2012). Given the relevance of CCN2 to interactions with integrins and the modulatory effect on chondrocytes further exploration of the relationship of this mechanism to de- and re-differentiation, and the development of organo-typic cultures is required.

7.4.5: Future work

Quantitative integration of data

To fully realize the potential of multiomics analysis a systematic integration methodology is required that also allows exploration and comparison of the different strata of quantitative omics data. Time constraints prevented exploration and implementation of these methods. As previously outlined a number of multivariate statistical integration methods have been utilised by researchers to integrate quantitative multi-omics data. Of these approaches projection-based methods have been widely used and are also available as open-source software packages in the *R* language. The `omnicade4` package implements a multiple co-inertia analysis (MCIA) approach to integrate greater than two omics data sets, a previous limitation of other methods (Meng, Kuster et al. 2014). This type of analysis was initially applied to ecology data to couple two (DolÉDec and Chessel 1994) or more (Bady, Dolédec et al. 2004) tables for the simultaneous ordination of data from a large number of environmental sites and populations. The MCIA method allows the projection of several data sets into the same dimensional space and on the same scale. This allows the exploration of diverse data features and the definition of co-relationships between usually highly dimensional omics data sets.

These methods still assume a common biological origin for the multi-omics data sets, i.e. the same tissue source or cell lines are assayed multiple times, and consequently this serves to limit analysis that may be undertaken with data available in this thesis. Given the narrow genetic background of the inbred rat strain used for the Affymetrix and proteomic data sets application of co-inertia analysis should be a valid quantitative integration strategy. Specifically this exploration would be used to

define co-relationships between the transcriptomic and proteomic data and assess the concordance between them. This is especially useful given than the proteomic analysis of cartilage and tendon data is not directly comparable. Given that this approach also doesn't require intersection of common feature annotations depth of coverage is retained potentially enhancing the power of pathway analysis. The potential for feature selection using MCIA also confers the possibility of an alternative *in silico* strategy for validation three-dimensional culture gene signatures.

Dealing with missing data

It is evident from the poor overlap of features across all three data sets that there is considerable loss of data points, which may bias and limit the understanding of the system. This is particularly an issue for proteomics data due to issues of sensitivity and range discussed in **Chapter 6**. Data points may be missing at random, however, many are missing as a result of non-random effects, e.g. qualitative changes between groups or annotation problems. Often in proteomics studies undetected proteins in surveys are assigned 'zero' values. Methods are required to impute or model this missing data (Aittokallio 2010, Li, Nie et al. 2011).

Basic statistical imputation was performed for missing qPCR data (section **2.2.7**), however, further exploration of missing data points for transcriptomic and proteomic profiling studies has not been undertaken. Exploration of the degree of loss with technical replicates would be useful, but the use of imputation methods may be considered as cost-efficient as additional replicates (Aittokallio 2010). Several approaches have been used for microarray data: *K*-nearest neighbours, Bayesian PCA, and least squares regression. Notably relatively simple imputation methods, such as using mean values, may work as well as more complex approaches (de Souto, Jaskowiak et al. 2015). At present these strategies require

implementation prior to downstream analysis in independent data sets. A pipeline for the imputation of missing data and subsequent omics integration should be developed to fully exploit the data arising from these studies.

7.4.6: Summary statements

The aim of this chapter was to integrate data from two transcriptomic and one proteomic data set with view to proposing a regulatory network to model three-dimensional organotypic culture systems and validate in future work. Clearly the postulated mechanistic networks do not include all elements of the system, but the objective was to deconstruct the mRNA and protein data represented in the different experiments, not to the point of abstraction, rather into a pragmatic mechanism that could be reasonably validated in the laboratory.

Two key themes emerge from an understanding of the components of the model networks presented in this chapter relating to the TGF-β signaling pathway and its involvement at the beginning and the end – in development and disease. This is derived from, firstly, the association with musculoskeletal developmental phenotypes (Havis, Bonnin et al. 2014, Lorda-Diez, Montero et al. 2014), the second with pro-fibrotic pathological mechanisms related to dysregulation of matrix turnover. For example TGF-β1, CCN2 and SERPINE1 are all co-expressed in a ureter-obstruction model of kidney fibrosis (Samarakoon, Dobberfuhl et al. 2013).

The contributory studies in this chapter consider whole tissue and matrix-depleted chondrocytes and tenocytes – these models do not consider the matrix roles in the sequestration and activation of TGF-β signaling. There is no supporting evidence regarding the temporal nature of this TGF-β signal (Cellière, Fengos et al. 2011). It is proposed that to define the prevailing nature of dedifferentiated musculoskeletal

cells, or their re-differentiation, requires further research aimed at collecting dynamic quantitative and systematic data with levels of down-stream targets such as CTGF, SMAD7 and SERPINE1 defined in parallel. In particular, development of cell-specific (Zi, Chapnick et al. 2012), and culture-specific, mathematical models for TGF-β signaling are required as the unified mechanism proposed assumes a common trajectory for both chondrocytes and tenocytes in culture.

To fully utilize the multi-level omics data available in this thesis data imputation and integration methods should be explored. Simple statistical strategies for imputation of missing data may be appropriate, whilst multivariate projection-based approaches should be employed to integrate quantitative data from gene and protein profiling studies.

References

Aittokallio, T. (2010). "Dealing with missing values in large-scale studies: microarray data imputation and beyond." Brief Bioinform **11**(2): 253-264.

Azuaje, F., H. Zheng, A. Camargo and H. Wang (2011). "Systems-based biological concordance and predictive reproducibility of gene set discovery methods in cardiovascular disease." Journal of Biomedical Informatics **44**(4): 637-647.

Bady, P., S. Dolédec, B. Dumont and J.-F. Fruget (2004). "Multiple co-inertia analysis: a tool for assessing synchrony in the temporal variability of aquatic communities." Comptes Rendus Biologies **327**(1): 29-36.

Bergeron, S., E. Lemieux, V. Durand, S. Cagnol, J. Carrier, J. Lussier, M. Boucher and N. Rivard (2010). "The serine protease inhibitor serpinE2 is a novel target of ERK signaling involved in human colorectal tumorigenesis." Molecular Cancer **9**(1): 271.

Blaney Davidson, E., E. Vitters, W. van den Berg and P. van der Kraan (2006). "TGF beta-induced cartilage repair is maintained but fibrosis is blocked in the presence of Smad7." Arthritis Research & Therapy **8**(3).

Blum, B., A. Roose, O. Barrandon, R. Maehr, A. Arvanites, L. Davidow, J. Davis, Q. Peterson, L. Rubin, D. Melton and H. Okano (2014). "Reversal of β cell de-differentiation by a small molecule inhibitor of the TGFβ pathway." eLife **3**: e02809.

Buchholz, M., A. Biebl, A. Neesse, M. Wagner, T. Iwamura, G. Leder, G. Adler and T. Gress (2003). "SERPINE2 (protease nexin I) promotes extracellular matrix production and local invasion of pancreatic tumors in vivo." Cancer Research **63**(16): 4945-4951.

Cabrero, D., I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Rudloff, A. Conesa and J. Tegner (2014). "Data integration in the era of omics: current and future challenges." BMC Systems Biology **8**(Suppl 2): I1.

Cellière, G., G. Fengos, M. Hervé and D. Iber (2011). "Plasticity of TGF-β signaling." BMC Systems Biology **5**.

de Souto, M. C., P. Jaskowiak and I. Costa (2015). "Impact of missing data imputation methods on gene expression clustering and classification." BMC Bioinformatics **16**(1).

Demeo, D., T. Mariani, C. Lange, S. Srisuma, A. Litonjua, J. Celedon, S. Lake, J. Reilly, H. Chapman, B. Mecham, K. Haley, J. Sylvia, D. Sparrow, A. Spira, J. Beane, V. Pinto-Plata, F. Speizer, S. Shapiro, S. Weiss and E. Silverman (2006). "The SERPINE2 gene is associated with chronic obstructive pulmonary disease." American Journal of Human Genetics **78**(2): 253-264.

DolÉDec, S. and D. Chessel (1994). "Co-inertia analysis: an alternative method for studying species–environment relationships." Freshwater Biology **31**(3): 277-294.

El-Gohary, Y., S. Tulachan, J. Wiersch, P. Guo, C. Welsh, K. Prasadan, J. Paredes, C. Shiota, X. Xiao, Y. Wada, M. Diaz and G. Gittes (2014). "A smad signaling network regulates islet cell proliferation." Diabetes **63**(1): 224-236.

Estrada, K., W. Wang, K. Retting, C. Chien, F. Elkhoury, R. Heuchel and K. Lyons (2013). "Smad7 regulates terminal maturation of chondrocytes in the growth plate." Developmental Biology **382**(2): 375-384.

Finnson, K., Y. Chi, G. Bou-Gharios, A. Leask and A. Philip (2012). "TGF-b signaling in cartilage homeostasis and osteoarthritis." Frontiers in Bioscience (Scholar edition) **4**: 251-268.

Gomez-Cabrero, D., I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa and J. Tegner (2014). "Data integration in the era of omics: current and future challenges." BMC Systems Biology **8**(Suppl 2): I1.

González, I., S. Déjean, P. Martin and A. Baccini (2008). "An R package to extend canonical correlation analysis." Journal of Statistical Software **23**(12): NA.

Grotendorst, G. R. (1997). "Connective tissue growth factor: a mediator of TGF-beta action on fibroblasts." Cytokine & Growth Factor Reviews **8**(3): 171-179.

Grotendorst, G. R., H. Okochi and N. Hayashi (1996). "A novel transforming growth factor beta response element controls the expression of the connective tissue growth factor gene." Cell Growth & Differentiation **7**(4): 469-480.

Günther, O., H. Shin, R. Ng, R. McMaster, B. McManus, P. Keown, S. Tebbutt and K.-A. Lê Cao (2014). "Novel multivariate methods for integration of genomics and proteomics data: applications in a kidney transplant rejection study." Omics : A Journal of Integrative Biology **18**(11): 682-695.

Guo, B., K. Inoki, M. Isono, H. Mori, K. Kanasaki, T. Sugimoto, S. Akiba, T. Sato, B. Yang, R. Kikkawa, A. Kashiwagi, M. Haneda and D. Koya (2005). "MAPK/AP-1-dependent regulation of PAI-1 gene expression by TGF-beta in rat mesangial cells." Kidney International **68**(3): 972-984.

Haider, S. and R. Pal (2013). "Integrated analysis of transcriptomic and proteomic data." Current Genomics **14**(2): 91-110.

Havis, E., M.-A. Bonnin, I. Olivera-Martinez, N. Nazaret, M. Ruggiu, J. Weibel, C. Durand, M.-J. Guerquin, C. Bonod-Bidaud, F. Ruggiero, R. Schweitzer and D. Duprez (2014). "Transcriptomic analysis of mouse limb tendon cells during development." Development **141**(19): 3683-3696.

Heit, C., B. Jackson, M. McAndrews, M. Wright, D. Thompson, G. Silverman, D. Nebert and V. Vasiliou (2013). "Update of the human and mouse SERPIN gene superfamily." Human Genomics **7**.

Hoshijima, M., T. Hattori, M. Inoue, D. Araki, H. Hanagata, A. Miyauchi and M. Takigawa (2006). "CT domain of CCN2/CTGF directly interacts with fibronectin and enhances cell adhesion of chondrocytes through integrin alpha5beta1." FEBS Letters **580**(5): 1376-1382.

Igarashi, A., H. Okochi, D. M. Bradham and G. R. Grotendorst (1993). "Regulation of connective tissue growth factor gene expression in human skin fibroblasts and during wound repair." Molecular Biology of the Cell **4**(6): 637-645.

Ito, Y., P. Bringas, A. Mogharei, J. Zhao, C. Deng and Y. Chai (2002). "Receptor-regulated and inhibitory Smads are critical in regulating transforming growth factor

β –mediated Meckel's cartilage development." Developmental dynamics : an official publication of the American Association of Anatomists **224**(1): 69-78.

Ivkovic, S., B. Yoon, S. Popoff, F. Safadi, D. Libuda, R. Stephenson, A. Daluiski and K. Lyons (2003). "Connective tissue growth factor coordinates chondrogenesis and angiogenesis during skeletal development." Development (Cambridge, England) **130**(12): 2779-2791.

Kavsak, P., R. K. Rasmussen, C. G. Causing, S. Bonni, H. Zhu, G. H. Thomsen and J. L. Wrana (2000). "Smad7 Binds to Smurf2 to Form an E3 Ubiquitin Ligase that Targets the TGF β Receptor for Degradation." Molecular Cell **6**(6): 1365-1375.

Kubota, S. and M. Takigawa (2015). "Cellular and molecular actions of CCN2/CTGF and its role under physiological and pathological conditions." Clinical Science (London, England : 1979) **128**(3): 181-196.

Kuo, T.-C., T.-F. Tian and Y. Tseng (2013). "3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data." BMC Systems Biology **7**(1): 64.

Lê Cao, K.-A., S. Boitard and P. Besse (2011). "Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems." BMC Bioinformatics **12**(1): 253.

Lê Cao, K.-A., I. González and S. Déjean (2009). "integrOmics: an R package to unravel relationships between two omics datasets." Bioinformatics (Oxford, England) **25**(21): 2855-2856.

Leask, A. and D. Abraham (2006). "All in the CCN family: essential matricellular signaling modulators emerge from the bunker." Journal of Cell Science **119**(Pt 23): 4803-4810.

Leask, A., S. Parapuram, X. Shi-Wen and D. J. Abraham (2009). "Connective tissue growth factor (CTGF, CCN2) gene regulation: a potent clinical bio-marker of fibroproliferative disease?" Journal of Cell Communication and Signaling **3**(2): 89-94.

Lee, C., B. Shah, E. Moioli and J. Mao (2010). "CTGF directs fibroblast differentiation from human mesenchymal stem/stromal cells and defines connective tissue healing in a rodent injury model." The Journal of Clinical Investigation **120**(9): 3340-3349.

Li, F., L. Nie, G. Wu, J. Qiao and W. Zhang (2011). "Prediction and Characterization of Missing Proteomic Data in Desulfovibrio vulgaris." Comparative and Functional Genomics **2011**(Article ID 780973): NA.

Liu, S.-C., C.-J. Hsu, H.-T. Chen, H.-K. Tsou, S.-M. Chuang and C.-H. Tang (2012). "CTGF increases IL-6 expression in human synovial fibroblasts through integrin-dependent signaling pathway." PloS ONE **7**(12): e51097.

Loeser, R. F. (2014). "Integrins and chondrocyte–matrix interactions in articular cartilage." Matrix Biology **39**: 11-16.

Lorda-Diez, C., J. Montero, J. Garcia-Porrero and J. Hurle (2014). "Divergent differentiation of skeletal progenitors into cartilage and tendon: lessons from the embryonic limb." ACS Chemical Biology **9**(1): 72-79.

Lorda-Diez, C., J. Montero, C. Martinez-Cue, J. Garcia-Porrero and J. Hurle (2009). "Transforming growth factors beta coordinate cartilage and tendon differentiation in the developing limb mesenchyme." The Journal of Biological Chemistry **284**(43): 29988-29996.

Madry, H. and M. Cucchiarini (2013). "Advances and challenges in gene-based approaches for osteoarthritis." The Journal of Gene Medicine **15**(10): 343-355.

Maeda-Uematsu, A., S. Kubota, H. Kawaki, K. Kawata, Y. Miyake, T. Hattori, T. Nishida, N. Moritani, K. Lyons, S. Iida and M. Takigawa (2014). "CCN2 as a novel molecule supporting energy metabolism of chondrocytes." Journal of Cellular Biochemistry **115**(5): 854-865.

Małgorzewicz, S., E. Skrzypczak-Jankun and J. Jankun (2013). "Plasminogen activator inhibitor-1 in kidney pathology (Review)." International Journal of Molecular Medicine **31**(3): 503-510.

Manoli, T., N. Gretz, H.-J. Gröne, M. Kenzelmann, R. Eils and B. Brors (2006). "Group testing for pathway analysis improves comparability of different microarray datasets." Bioinformatics (Oxford, England) **22**(20): 2500-2506.

Massagué, J. (2012). "TGFβ signalling in context." Nature Reviews. Molecular Cell Biology **13**(10): 616-630.

Massagué, J., J. Seoane and D. Wotton (2005). "Smad transcription factors." Genes & Development **19**(23): 2783-2810.

McKee, C., D. Xu, Y. Cao, S. Kabraji, D. Allen, V. Kersemans, J. Beech, S. Smart, F. Hamdy, A. Ishkanian, J. Sykes, M. Pintile, M. Milosevic, T. van der Kwast, G. Zafarana, V. R. Ramnarine, I. Jurisica, C. Mallof, W. Lam, R. Bristow and R. Muschel (2012). "Protease nexin 1 inhibits hedgehog signaling in prostate adenocarcinoma." The Journal of Clinical Investigation **122**(11): 4025-4036.

Meng, C., B. Kuster, A. n. Culhane and A. Gholami (2014). "A multivariate approach to the integration of multi-omics datasets." BMC Bioinformatics **15**(1): 162.

Morales, M. G., J. Gutierrez, C. Cabello-Verrugio, D. Cabrera, K. Lipson, R. Goldschmeding and E. Brandan (2013). "Reducing CTGF/CCN2 slows down mdx muscle dystrophy and improves cell therapy." Human Molecular Genetics **22**(24): 4938-4951.

Nakanishi, T., T. Nishida, T. Shimo, K. Kobayashi, T. Kubo, T. Tamatani, K. Tezuka and M. Takigawa (2000). "Effects of CTGF/Hcs24, a product of a hypertrophic chondrocyte-specific gene, on the proliferation and differentiation of chondrocytes in culture." Endocrinology **141**(1): 264-273.

Nakao, A., M. Afrakhte, A. Morén, T. Nakayama, J. L. Christian, R. Heuchel, S. Itoh, M. Kawabata, N. E. Heldin, C. H. Heldin and P. ten Dijke (1997). "Identification of Smad7, a TGFbeta-inducible antagonist of TGF-beta signalling." Nature **389**(6651): 631-635.

Oemar, B. S. and T. F. Lüscher (1997). "Connective tissue growth factor. Friend or foe?" Arteriosclerosis, Thrombosis, and Vascular Biology **17**(8): 1483-1489.

Omoto, S., K. Nishida, Y. Yamaai, M. Shibahara, T. Nishida, T. Doi, H. Asahara, T. Nakanishi, H. Inoue and M. Takigawa (2004). "Expression and localization of connective tissue growth factor (CTGF/Hcs24/CCN2) in osteoarthritic cartilage." Osteoarthritis and Cartilage / OARS, Osteoarthritis Research Society **12**(10): 771-778.

Palsson, B. and K. Zengler (2010). "The challenges of integrating multi-omic data sets." Nat Chem Biol **6**(11): 787-789.

Qi, W., X. Chen, S. Twigg, Y. Zhang, R. Gilbert, D. Kelly and C. Pollock (2007). "The differential regulation of Smad7 in kidney tubule cells by connective tissue growth factor and transforming growth factor-beta1." Nephrology (Carlton, Vic.) **12**(3): 267-274.

Sakou, T., T. Onishi, T. Yamamoto, T. Nagamine, K. Sampath and P. ten Dijke (1999). "Localization of Smads, the TGF-β Family Intracellular Signaling Components During Endochondral Ossification." Journal of Bone and Mineral Research **14**(7): 1145-1152.

Samarakoon, R., A. Dobberfuhl, C. Cooley, J. Overstreet, S. Patel, R. Goldschmeding, K. Meldrum and P. Higgins (2013). "Induction of renal fibrotic genes by TGF-β1 requires EGFR activation, p53 and reactive oxygen species." Cellular Signalling **25**(11): 2198-2209.

Samarakoon, R., J. Overstreet and P. Higgins (2013). "TGF-β signaling in tissue fibrosis: redox controls, target genes and therapeutic opportunities." Cellular Signalling **25**(1): 264-268.

Sappino, A. P., R. Madani, J. Huarte, D. Belin, J. Z. Kiss, A. Wohlwend and J. D. Vassalli (1993). "Extracellular proteolysis in the adult murine brain." The Journal of Clinical Investigation **92**(2): 679-685.

Sass, S., F. Buettner, N. Mueller and F. Theis (2013). "A modular framework for gene set analysis integrating multilevel omics data." Nucleic Acids Research **41**(21): 9622-9633.

Sass, S., F. Buettner, N. Mueller and F. Theis (2014). "RAMONA: a Web application for gene set analysis on multilevel omics data." Bioinformatics (Oxford, England).

Scharstuhl, A., E. Vitters, P. van der Kraan and W. van den Berg (2003). "Reduction of osteophyte formation and synovial thickening by adenoviral overexpression of transforming growth factor beta/bone morphogenetic protein inhibitors during experimental osteoarthritis." Arthritis and Rheumatism **48**(12): 3442-3451.

Schmierer, B. and C. Hill (2007). "TGF[beta]-SMAD signal transduction: molecular specificity and functional flexibility." Nat Rev Mol Cell Biol **8**(12): 970-982.

Schweitzer, R., E. Zelzer and T. Volk (2010). "Connecting muscles to tendons: tendons and musculoskeletal development in flies and vertebrates." Development (Cambridge, England) **137**(17): 2807-2817.

Sobral, L., P. F. Montan, K. G. Zecchin, H. Martelli-Junior, P. A. Vargas, E. Graner and R. Coletta (2011). "Smad7 blocks transforming growth factor-β 1-induced gingival fibroblast-myofibroblast transition via inhibitory regulation of Smad2 and connective tissue growth factor." Journal of Periodontology **82**(4): 642-651.

Takigawa, M. (2013). "CCN2: a master regulator of the genesis of bone and cartilage." Journal of Cell Communication and Signaling **7**(3): 191-201.

Tchetina, E. (2011). "Developmental mechanisms in articular cartilage degradation in osteoarthritis." Arthritis **2011**: Article ID: 683970.

Tomita, N., T. Hattori, S. Itoh, E. Aoyama, M. Yao, T. Yamashiro and M. Takigawa (2013). "Cartilage–Specific Over-Expression of CCN Family Member 2/Connective Tissue Growth Factor (CCN2/CTGF) Stimulates Insulin-Like Growth Factor Expression and Bone Growth." PLoS ONE **8**(3): e59226.

Vaillant, C., O. Michos, S. Orolicki, F. Brellier, S. Taieb, E. Moreno, H. Té, R. Zeller and D. Monard (2007). "Protease nexin 1 and its receptor LRP modulate SHH signalling during cerebellar development." Development (Cambridge, England) **134**(9): 1745-1754.

van Rooyen, B., G. Schäfer, V. Leaner and I. Parker (2013). "Tumour cells down-regulate CCN2 gene expression in co-cultured fibroblasts in a Smad7- and ERK-dependent manner." Cell Communication and Signaling : CCS **11**.

Wang, K., B. Wang, A. Y. Xing, K. S. Xu, G. X. Li and Z. H. Yu (2014). "Prognostic significance of SERPINE2 in gastric cancer and its biological function in SGC7901 cells." Journal of Cancer Research and Clinical Oncology.

Wang, Z., Y. Qiu, J. Lu and N. Wu (2013). "Connective tissue growth factor promotes interleukin-1β-mediated synovial inflammation in knee osteoarthritis." Molecular Medicine Reports **8**(3): 877-882.

Wilson, R., D. Belluoccio, C. Little, A. Fosang and J. Bateman (2008). "Proteomic characterization of mouse cartilage degradation in vitro." Arthritis and Rheumatism **58**(10): 3120-3131.

Wilson, R., J. Whitelock and J. Bateman (2009). "Proteomics makes progress in cartilage and arthritis research." Matrix Biology **28**(3): 121-128.

Xiao, X., I. Gaffar, P. Guo, J. Wiersch, S. Fischbach, L. Peirish, Z. Song, Y. El-Gohary, K. Prasadan, C. Shiota and G. Gittes (2014). "M2 macrophages promote beta-cell proliferation by up-regulation of SMAD7." Proceedings of the National Academy of Sciences of the United States of America **111**(13).

Yano, M., Y. Inoue, T. Tobimatsu, G. Hendy, L. Canaff, T. Sugimoto, S. Seino and H. Kaji (2012). "Smad7 inhibits differentiation and mineralization of mouse osteoblastic cells." Endocrine Journal **59**(8): 653-662.

Yao, F., J. Coquery and K.-A. Cao (2012). "Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets." BMC Bioinformatics **13**(1): 24.

Zhang, Y., X. H. Feng and R. Derynck (1998). "Smad3 and Smad4 cooperate with c-Jun/c-Fos to mediate TGF-beta-induced transcription." Nature **394**(6696): 909-913.

Zhen, G. and X. Cao (2014). "Targeting TGFβ signaling in subchondral bone and articular cartilage homeostasis." Trends in Pharmacological Sciences **35**(5): 227-236.

Zhen, G., C. Wen, X. Jia, Y. Li, J. Crane, S. Mears, F. Askin, F. Frassica, W. Chang, J. Yao, J. Carrino, A. Cosgarea, D. Artemov, Q. Chen, Z. Zhao, X. Zhou, L. Riley, P. Sponseller, M. Wan, W. W. Lu and X. Cao (2013). "Inhibition of TGF-β signaling in mesenchymal stem cells of subchondral bone attenuates osteoarthritis." Nature Medicine **19**(6): 704-712.

Zhu, G., L. Warren, J. Aponte, A. Gulsvik, P. Bakke, W. Anderson, D. Lomas, E. Silverman and S. Pillai (2007). "The SERPINE2 gene is associated with chronic obstructive pulmonary disease in two large populations." American Journal of Respiratory and Critical Care Medicine **176**(2): 167-173.

Zhu, L., S. Chen and Y. Chen (2011). "Unraveling the biological functions of Smad7 with mouse models." Cell & Bioscience **1**.

Zi, Z., D. A. Chapnick and X. Liu (2012). "Dynamics of TGF-β /Smad signaling." FEBS Letters **586**(14): 1921-1928.

8: General Discussion

8.1: Project objectives revisited

Statements of intent

In this thesis a systems biology approach was employed to explore the three-dimensional culture phenotype and relate this to whole tissue and standard monolayer culture. As defined at the outset monolayer culture has existed as the pre-eminent approach for *in vitro* modeling for cartilage and tendon. The limitations of this strategy have been clear for many years and novel organotypic culture systems have been devised to overcome these limitations. Despite this, a paradigm shift from two- to three-dimensional culture systems has not occurred. At the time of writing no systematic investigation of three-dimensional models of cartilage or tendon has been undertaken. Furthermore, these model culture methods are commonplace within the field of musculoskeletal research in spite of the fact that the gene and protein profiles of these models have not been fully characterized.

It was asserted in the introduction that without validated and standardised organotypic models expedient translation of bioengineered tissue from the laboratory to the clinics would not be possible within the current regulatory frameworks. It was also stated that the majority of studies utilized narrow definitions of differentiated status when qualifying progress in cartilage and tendon tissue engineering. To fully realize the potential of *in vitro* systems, maximize the use of rare clinical specimens, and reduce the use of animal models there is a need for global characterization of organo-typic models and systematic comparisons with the *in vivo* environment. Systems biology methods,

such as those presented in this thesis, are well-suited to complex integration of data and may be used to generate predictive models for functional validation.

Associated with each chapter specific observations have been presented and specific discussion statements built upon these; associated recommendations for further validation and development of the results were also defined. In this final discussion the principal themes of the study as a whole are considered; additionally the contribution to musculoskeletal research is outlined and a road map for further work is provided.

The key project objectives were defined in **Chapter 1** and these are revisited here in three sections. Novels findings and their impact are reflected upon relative to these objectives; general concerns arising across all studies are considered, and future research and methodologies are presented.

8.2: The de- and re-differentiated phenotypes.

Objective 1: To define dedifferentiation and re-differentiation to mark-out the phenotypic boundaries within which cartilage and tendon cells function

The terms de- and re-differentiation are used in cartilage and tendon research with no mechanistic definition. Additionally there is no standard phenotype against which to benchmark novel findings and progress in bioengineering. Progress is defined through the analysis of a limited spectrum of well-described differentially expressed genes and proteins, but inter-experimental comparisons are difficult and rarely are findings interrogated against other studies in the field. By creating a reference gene expression profile for native cartilage and tendon, dedifferentiated chondrocytes and tenocytes, and three-dimensional culture systems a baseline catalogue is created against which further developments may be referred.

Regenerative medicine has the goal of replacing lost or damaged cells. Approaches to achieve this include reprogramming, dedifferentiation or transdifferentiation; of these dedifferentiation represents a proliferative, less differentiated phenotype (Jopling, Boue et al. 2011). In the mammal, recent work has found dedifferentiation occurs in regenerative (Porrello, Mahmoud et al. 2011) and degenerative contexts (Szibor, Pöling et al. 2014). It may be a prerequisite to the replacement of lost cells (El-Gohary, Tulachan et al. 2014). Dedifferentiation is recognised in chondrocytes (Schulze-Tanzil 2009) and tenocytes (Yao, Bestwick et al. 2006) as a 'side-effect' of monolayer culture. Dedifferentiation as a mechanism in osteoarthritis is inconsistently referred to (Young, Smith et al. 2005) and its contribution to tendinopathy is unknown. In order to investigate dedifferentiation as a contributory process to a loss of function or degenerative phenotype, or even a regenerative mechanism, in adult cartilage and tendon there must be adequate reference points.

8.2.1: Dedifferentiation in monolayer represents a proliferative 'pre-differentiated' phenotype

To initially explore a sematic association with dedifferentiation in regenerative contexts (Kragl, Knapp et al. 2009) qualitative definitions of dedifferentiation in monolayer were made: i) quiescent cells proliferate, ii) there is loss or reduced expression of the functional synthetic profile, iii) increased expression of markers of a pre-differentiated phenotype, iv) and markers are lineage restricted and not indicative of reversion to a pluripotent state, i.e. embryonic stem cell.

Across two gene expression profiling studies there was evidence that chondrocytes and tenocytes were proliferative and expression of functional hallmarks were reduced (*Tnmd*, *Col2a1*) (**2.3.3** and **3.3.3**) in line with current understanding. This study contributes to this by demonstrating increases in the expression of mesenchymal markers and markers

of cartilage and tendon development (including *Scx* and *Mkx*) in monolayer-expanded cells. There was no evidence for a reversion to a pluripotent state as defined by the expression of markers associated with embryonic or induced pluripotent stem cells. Furthermore, the higher expression of homeobox genes, associated with developmental positioning (*Pitx1*), was robustly expressed in monolayer. Finally, evidence of preserved expression of homeobox genes associated with topographical anatomy was described for adult cartilage and tendon (section **2.3.9**).

In purely sematic terms there is reasonable concordance with the qualitative definitions. This thesis proposes that chondrocytes and tenocytes in monolayer culture have a phenotype consistent with a proliferative 'pre-differentiated' musculoskeletal cell that may have position- and lineage-restricted potential. In other words, the simple act of placing cells in monolayer culture induces phenotypic changes reminiscent of pre-differentiated state. This notion challenges studies defining the presence of adult tissue progenitor cells based exclusively on the expression of marker genes in monolayer.

Recent evidence suggests that cartilage and tendon precursor cells are not defined by the isolated expression of *Sox9* or *Scx* alone, rather combinations of expression patterns are evident in precursor tenocytes depending on the developmental proximity to nascent cartilage (Sugimoto, Takimoto et al. 2013) (section **1.5**). Evidence of the expression of *Scx* in both chondrocytes and tenocytes at passage three (section **3.3.3**) lends credence to the hypothesis that they show common heritage and respond in a comparable way *in vitro*. This has implications for our understanding of chondrocyte and tenocyte cell biology and how we exploit these cells in tissue engineering. Firstly, these adult-derived, terminally differentiated cells are plastic and have potential to express developmental markers. Secondly, by expressing developmental markers dedifferentiated cells may be in a permissive state that may be manipulated for regenerative purposes through the

correct biochemical and biomechanical cues. Thirdly, this study provides reference markers against which gene expression profiles from disease states may be compared to in order to make inferences about underlying pathophysiological mechanisms, i.e. is loss of function by dedifferentiation a mechanism in chronic cartilage degeneration?

8.2.2: Tissue-derived stem cells in monolayer cultures: dedifferentiated cells by any other name?

Evidence of the multi-potentiality of tissue-derived stem cells has long relied on qualitative colourimetric assays (Alcian blue and Oil Red O staining for chondrogenic and adipogenic differentiation potential respectively) and limited panels of genes are assessed by qPCR. There is no published comparison between the transcriptome and proteome of putative tissue-derived stem cells that have undergone directed differentiation studies and mature tissue. Methods for isolating tissue progenitor populations from tendon (Bi, Ehirchiou et al. 2007), for example, are not dissimilar to those used in this thesis to derive tenocytes for monolayer culture. In both scenarios cultures start with very low seeding densities. Furthermore, there is no standardization in these techniques or comparison of putative tissue-progenitor cells between laboratories (Prockop 2009).

The markers for musculoskeletal progenitors were frequently evident in the gene expression surveys in this thesis. For example, Worthley, *et al* (2015) recently defined a novel nestin-negative (BMSC marker) skeletal stem cell population, osteochondroreticular (OCR) cells, in a cell-tracking experiment based upon the expression of the BMP antagonist *Grem1* (Gremlin 1) (Worthley, Churchill et al. 2015). Consistently *Grem1* has been identified as significantly differentially expressed in three-dimensional culture systems from both chondrocytes and tenocytes (sections **2.3.2**, **3.3.2**, **6.3.3**, and **SD2/3**). This may indicate that common regulatory mechanisms that

are in play in the de- to re-differentiation transition are common to differentiation pathways for resident adult progenitor cells.

Markers for musculoskeletal progenitor cells are diverse and inconsistent in the literature and are over-represented in differential gene expression analysis of proliferating chondrocytes and tenocytes in this thesis. A critical area for development of the findings in this study would be global gene and protein profiling of mesenchymal stem cells relative to cartilage stem/progenitor cells (Jiang and Tuan 2015), cells undergoing directed differentiation towards a chondrocytic phenotype, three-dimensional culture systems and monolayer (dedifferentiated) cells. Without more subtle descriptors the phenotypes of musculoskeletal cells in diverse environmental conditions cannot be adequately resolved and concerns should be raised where adult stem cell markers are reported to be present in cells that have undergone periods of monolayer expansion.

8.2.3: Gene expression convergence challenges continued use of monolayer to model chondrocyte and tenocyte phenotypes

Convergence of gene expression profiles in monolayer culture has been shown in cells from multiple tissue sources (Prasad, Kumar et al. 2013). In two independent comparative studies convergence of chondrocytes and tenocyte gene expression profiles were found (sections **2.3.2** and **3.3.2**). The fundamental question arising from this finding is whether the use of monolayer culture is defensible for the study of chondrocytes or tenocytes given that monolayer culture is physiologically uninformative (Haycock 2011). Clearly there are limitations to the use of whole tissue, but universal acceptance of three-dimensional culture systems, and a move away from monolayer culture studies, is restrained by a lack of commercially available 'bio-inspired' materials that could be used in a flexible, reproducible and affordable manner to facilitate

seamless *in vitro* to *in vivo* translation of research (Prestwich 2007). Despite the importance of monolayer culture this study identifies that a critical obstacle to development of rational organotypic models is a dearth of integrated mRNA expression and protein abundance studies and a reliance on underpowered studies defining changes in a limited number of established markers as the sole deliverable output of a study. A deeper understanding of the nature of organotypic cultures and the protein regulatory networks that govern the phenotype would facilitate translational research. The contributions to this end are detailed further in the succeeding sections.

8.2.4: Three-dimensional cultures do not restitute native expression profiles

For three-dimensional cultures to represent adequate models of tissue a stable phenotype is required with gene and protein profiles consistent with (or comparable to) the tissue under investigation. To achieve this would entail integration of biophysical and biochemical signals dynamically (Spanoudes, Gaspar et al. 2014). In contribution to this goal mRNA and protein profiling evidence in this thesis demonstrates that established three-dimensional culture models for chondrocytes (De Ceuninck, Lesur et al. 2004) and tenocytes (Kapacee, Richardson et al. 2008) fail to restitute dedifferentiated cells to a cartilage or tendon phenotype respectively.

For alginate culture, in particular, there is evidence of elevated expression of cytokines (sections **2.3.2** and **3.3.2**), a shift to oxidative phosphorylation (section **6.3.5**) and evidence of reduced cell viability with time (section **2.3.8**). Alginate cultures in both rat and human gene expression studies had more extreme expression of markers of inflammation than cartilage samples derived from human osteoarthritis or rat surgical models of osteoarthritis (section **5.3.9**).

8.2.5: Gene signature classifies three-dimensional cultures

Implicit in the title of this study was the requirement that the analysis of large expression profiling studies would have a practical output in relation to tissue engineering rather than creation of reference gene and protein lists alone. A thirteen gene signature with enriched protein-protein interactions confidently identified alginate cultures separate from native or monolayer samples (section **5.3.10**). The results of the microarray meta-analysis may facilitate a systems approach to organotypic model development by defining whether gene expression changes associated with a novel intervention, e.g. novel cell scaffold, are more comparable to a standardised model systems or a native phenotype. The validation of a standard phenotype in these culture systems would establish a control condition against which these interventions could be compared. It could also encourage the reduction of animal tissue harvesting for these studies. This study also established that the gene signature differed between human and rodent studies with possible implications related to the translation of findings in rodent musculoskeletal studies to human disease.

8.2.6: Matrix-depletion facilitates deeper exploration of cartilage and tendon proteome

Improving the depth of proteome coverage usually requires additional fractionation or depletion steps to deal with the vast dynamic range in protein abundance. High resolution LC-MS has improved this in recent years, however, this is still challenging for cartilage and tendon due to the highly anionic extra-cellular matrix. In this thesis a simple protocol is outlined to deplete chondrocytes and tenocytes of extra- and peri-cellular matrix resulting in a greater depth of coverage than found previous discovery publications (Hsueh, Önnerfjord et al. 2014), (section **6.3.2**). This opens the potential of detecting low abundance, regulatory proteins – for example, gremlin 1, catenin beta 1 and MAP kinases were all identified in tenocytes in monolayer or fibrin cultures.

Beyond validation of individual targets the natural progression for this analysis would be to systematically build-up the proteome from component studies. In Deshmuhk, *et al* (2015) comparable problems associated with skeletal muscle discovery analysis were overcome by combining uniquely identified proteins from C2C12 cultured myotubes and whole muscle to define 10, 218 proteins (Deshmukh, Murgia et al. 2015). Stepwise discovery analysis of fractionated and cultured samples, in addition to whole tissue, would aid a more comprehensive understanding of the cartilage and tendon proteome.

8.3: Dealing with complex and heterogeneous data sets

Objective 2: Define cross-species responses to homeostatic perturbations by cartilage and tendon through integration of gene-expression data

8.3.1: Cross-species comparison of transcriptome networks yields novel insights

In order to challenge the conclusions drawn from gene expression surveys of de- and re-differentiated cell profiles were considered with respect to other published data sets. The dearth of quality microarray data sets profiling cartilage or tendon in the rat required a permissive study inclusion policy in order to apply co-expression analysis methods to explore the global transcriptome network. Using a data merging approach resulted in a complex and heterogenous matrix of samples and expression data. The same methods were applied to human cartilage and tendon data and a comparison of the global transcriptome network architecture was undertaken to define common functional sub-networks that had a strong statistical association with particular phenotypes.

Modules within the rat transcriptome network corroborated the gene ontology functional annotations described in **Chapter 2** and **3** with gene modules from 170 arrays describing immune system processes, cell cycle and metabolic activity, and skeletal system development. From the rat network a highly interconnected sub-network of genes with known osteoarthritis associations were found (section **5.3.7**). These were strongly associated with a normal cartilage phenotype and strongly negatively correlated with perturbed samples from rat models of osteoarthritis. Within this sub-network genes with no known osteoarthritis association were found – it is proposed that these genes may have relevance to a further understanding of the pathophysiology of osteoarthritis in model species. The impact on our understanding of human disease is discussed below.

8.3.2: Informing complex human disease from rodent models – thesis findings have practical relevance to researchers

The ultimate deliverable output of this study should be findings that allow some inference to be made with regard to human musculoskeletal disease – either to explore further regulatory networks that may be perturbed in the disease state, to refine disease models, and encourage development of regenerative strategies. As has been asserted in earlier statements rat *in vitro* models of cartilage and tendon homeostasis are physiologically imperfect and so it becomes difficult to untangle what may be usefully extrapolated to human research, what is species-specific, and what is artefact. In this thesis it is demonstrated that, using the available gene expression data sets for rat and human, there are co-expression network modules associated with normal cartilage gene expression profiles in the rat that are not replicated in human studies (section **5.3.7**). The reasons for this have been described elsewhere, but it highlights a core issue of a network approach using data from rodent *in vivo* and *in vitro* models. The lack of normal expression profiles for human cartilage and tendon make it difficult to compare

transcriptome networks; the absence of evidence for a given gene module in a co-expression network from human data is not evidence for its absence.

There is considerable controversy centred on the relevance of rodent models to human disease, most notably in inflammation (Seok, Warren et al. 2013, Takao and Miyakawa 2015). Primarily these studies have focused on the correlation of gene expression profiles between mice and human expression studies. In this thesis low, but positive, correlations of gene expression rank were found between species (section **5.3.4**) for cartilage and tendon derived studies. It is also demonstrated that the overall transcriptomic structure is highly conserved between rat and human networks (section **5.3.5**). This is consistent with a recent re-evaluation of mouse ENCODE data demonstrating that gene expression across species (mouse and human) clusters by tissue and not by species (Gilad and Mizrahi-Man 2015). The converse had been published by the ENCODE consortium (Lin, Lin et al. 2014), but supporting earlier statements in this thesis calling for disclosure of microarray sample batching in public repositories, batch effects in the ENCODE data had not been resolved.

Findings presented in this thesis have a practical impact for musculoskeletal researchers interested in developing bioengineered cartilage and tendon. For future studies to make further headway efforts should be made to at a community level to rationalize models, optimise laboratory methods/parameters, and curate gene orthologs across rodent and human studies not least because of the wider impact on the veracity of translational research in science policy and knock-on effects on funding. Further development of the co-expression analysis study presented in this thesis should bear the controversies, discussed above, in mind. For example, rodent and human expression profiles both consider inflammatory osteoarthritis, but activation of such pathways occur on vastly different time-scales. More careful matching of expression studies across species, and

dealing with batch effects where possible, is more likely to yield informative results than loose amalgams of tissue, cell and disease profiles (Shay, Lederer et al. 2015).

8.3.3: Ontologies fail to provide depth of functional annotation

The use of gene ontology functional annotations was a core data-mining tool in the collection of studies presented within this thesis. The presence of highly enriched annotations for immune functions in native cartilage, and muscle transcripts in tendon, from differentially expressed genes prompted exploration of methods to determine how these profiles compared to those from other studies. The possible methodological and anatomical reasons for this have been discussed in **Chapter 2**. Further consideration of the relevance of gene ontology annotations to a systems understanding is warranted. Throughout this thesis annotations using developmental terminology are evident: 'anatomical structure morphogenesis', 'tissue morphogenesis', 'regulation of cell differentiation' (sections **2.3.4**, **3.3.4**, and **7.3.1**) are found to be associated with dedifferentiation in monolayer and re-differentiation in three-dimensional culture systems. Enriched terms such as these contribute positively to a hypothesis of constrained plasticity of isolated chondrocytes and tenocytes. However, alongside these numerous incongruous annotations are evident; the most obvious example of this was the significant number of annotations associated with neurodegenerative disorders used to annotate enriched KEGG pathways in the proteomic profiles of monolayer and three-dimensional cultures (section **6.3.3**). A further example was the annotation of dedifferentiated chondrocytes with 'tendon development' annotation (section **3.3.4**), but this was not apparent for the equivalent tenocyte analysis. Disparity across knowledge bases was also evident with PI-3K signalling pathway shown to be the predominant pathway in de- and re-differentiation (section **3.3.7**) yet described as hepatic stellate cell activation by Ingenuity Pathway Analysis. With respect to the cross-species analysis

presented in this thesis assumptions are also made with regard to the equivalence of annotations.

These concerns with ontologies in musculoskeletal research may be considered as arising from several broad issues: i) heterogenous cell populations (especially for native tissues) obfuscate accurate annotation of the cell population of interest, ii) functional annotations are inconsistent across knowledge bases, iii) annotations are biased and are not stage-, condition-, or tissue-specific, iv) and annotations are subject to researcher bias.

Nehrt, *et al* (2011) used gene ontology annotations to refute the ortholog conjecture, that orthologous genes share greater functional similarity that paralogous genes (Nehrt, Clark et al. 2011). Although this has now been challenged within the context of comparative genomics it revealed issues within the gene ontological annotations, in particular the 'open world assumption' that the absence of a GO functional annotation for a gene does not indicate absence of that function, rather it reflects artifacts created by the methods of collection of molecular biology data. This has been described as a 'global ascertainment bias' in that certain types of experiments tend to be performed in certain model species resulting in very unequal representation of certain biological and molecular functions between rodent and human annotations. In particular biological process terms for development and cell differentiation are over-represented in the mouse relative to human annotations (Thomas, Wood et al. 2012). These issues are compounded in tissues for which there are few gene ontology annotations, i.e. tendon. Consequently, an understanding of how GO annotations are derived should inform the corroborative process for the data presented in this thesis, in particular comparative profiling of development stages of cartilage and tendon with de- and re-differentiated cells should be considered.

8.3.4: Deconvolution of data from heterogenous samples may yield further insights into cartilage and tendon

Gene expression profiles derived from cartilage and tendon represent the average signal for a heterogenous tissue. The relative proportions of each cell type can vary considerably, especially when derived from small tissue samples collected by manual dissection. This can confound downstream analysis, not only through functional annotations (described above), but also through the misidentification of candidate biomarkers.

To disentangle the expression signals requires the use of deconvolution algorithms; these can serve as cost-effective tools to rationalise the component cell profiles, which would otherwise require specialist resources, e.g., micro-dissection, cell sorting. Flexible frameworks, such as the `CellMix` *R* package (Gaujoux and Seoighe 2013) facilitate the implementation of these methods, but require auxiliary data sets to define cell proportions. Only recently a novel implementation of latent variable analysis in R, `CellCODE`, has been shown to be able to assign differentially expressed genes to cell-types based upon data structure alone and without specific knowledge of the data set (Chikina, Zaslavsky et al. 2015). Further analysis of expression data from whole cartilage and tendon would benefit from the application of deconvolution strategies to retrieve cell-specific profiles.

8.4: Challenges of data integration

Objective 3: Integrate gene expression and protein abundance data to rationalize validation targets and derive mechanistic networks.

8.4.1: Summary of integration approaches employed

The data accumulated in this collection of studies represents a rich resource to be exploited further. However, discussion of each study in isolation has limited relevance to a wider systems understanding of the response of chondrocytes and tenocytes to environmental perturbations. Through integration of data from snapshot profiles at different levels of the biological hierarchy a consistent and comprehensive narrative may emerge.

A number of data integration strategies were employed in this thesis. To aggregate data from multiple diverse microarray gene expression profiles data was merged by intersecting on common gene identifiers and applying a global normalization method to create a single matrix. The structure of the transcriptome network could then be determined through the use of a weighted gene co-expression network analysis (sections **4.3.1** and **5.3.3**). The utility of this approach was clear, however, small data sets demonstrated unstable network structures and in this context the confidence with which regulatory hub genes could be identified was marginal; this was exemplified by the equivocal impact of siRNA knock-down of the *Lzts2* gene on expression of markers of differentiation status in chondrocytes (section **4.3.7**). Despite the strong statistical associations supporting the choice of module and hub gene in an individual data set this study highlighted the necessity for interrogation of multiple data sets to define consensus modules.

To determine a comprehensive functional annotation of the biological processes likely to be involved in de- and re-differentiation a novel model-based Bayesian method was used (Sass, Buettner et al. 2014). This approach considers each data set as a noisy representation of a common underlying gene response. This permitted the integration of functional annotations from a number of biological levels, simultaneously dealing with redundant terms and issues of multiple testing. This application did summarise dedifferentiation by the terms 'ossification', 'oxidation-reduction process' and the KEGG Wnt-signalling pathway. The limitation of this approach, in addition to ontology concerns discussed above, is that only two different levels may be considered and makes no use of quantitative data.

The third method used to integrate data from transcriptomic and proteomic data sets aimed to infer the common upstream transcriptional regulators that resulted in the observed expression profiles from each data set. This analysis made use of causal analytic algorithms included within Ingenuity Pathway Analysis. Upstream regulators are predicted that are consistent with regulation of the observed gene expression profile through direct or indirect relationships. Mechanistic networks are hypothesis networks built upon identified upstream regulators by connecting the regulators considered to be acting through the same signalling mechanism (Krämer, Green et al. 2014).

This technique demonstrated the reciprocal activation and inhibition predicted for the master regulators TGF-β1 and PDGF BB in de- and re-differentiation. Furthermore, a highly enriched protein-protein interaction network was evident for dedifferentiation derived from Ingenuity mechanistic hypothesis networks. Through this SMAD7 and CTGF were identified as common downstream targets in the dedifferentiated phenotype (section **7.3.2**). For redifferentiation, mechanistic networks were centred on JUN. FOS, BMP2, GREM1 and IL-1B and IL-6. By performing this analysis over

several studies and two tissue sources it was possible to define a consensus network. This meets the goal of devising a mechanistic network for de- and re-differentiation and sub-setting a target group for further wet-lab validation and input to models.

Although this method allowed qualitative comparisons to be made across independent data sets, e.g. TGF-β1 activation is predicted in monolayer in all data sets (section **7.3.2**), this approach again does not utilize the quantitative data. Missing nodes in a regulatory cascade are given statistical predictions for their activation or inhibition. This allowed the inclusion of targets not shown to be differentially expressed, but were relevant to the tissue under investigation.

Clearly the observed gene expression profile may be modulated by a number of different regulators and it cannot be known *a priori* which of these predominates. Although each of these hypothesis networks has a statistical score associated with it they are a function of the observed gene expression profile. The absence of nodes and the strict cut-off defined in Ingenuity could have implications on the reproducibility of these networks.

8.4.2: Pathways and mechanistic networks

This study reports the contribution of a number of signalling pathways as being involved monolayer and three-dimensional culture phenotypes including Wnt-, NF-kB, PI-3K/Akt (section **3.3.7**), PPAR (section **6.3.6**), IL-6 (section **5.3.8**) and TGF-β (section **7.3.2**)-signalling. These have all been implicated in development and/or disease mechanisms in chondrocyte and tenocyte or defined by *in vitro* studies. The finding of multiple signalling pathways in this study is not contradictory rather it demonstrates the requirement of integration across multiple data sets and the levels of the biological hierarchy.

By defining the consensus upstream regulators associated with mRNA and protein profiles in this study it was possible to propose a unified mechanistic network for the observed chondrocyte and tenocyte phenotypes. Crucially the proposed network is static and incomplete; it requires both an assessment of sufficiency (whether it describes the behavior of the system within a specified range of tolerance) and a test of realism (whether this is the correct mechanism) (Boogerd, Bruggeman et al. 2013). In line with current opinion development of this TGF-β mediated model would require a more quantitative and systematic approach, not least mathematical modeling of temporal signals (Zi, Chapnick et al. 2012). Time course analysis would be an integral component of further expression profiling.

The future challenge from these, and other, expression data sets is the reverse engineering of gene regulatory networks (GRN) from gene expression data. In using co-expression network analysis functional sub-networks were isolated and these are useful for exploring emerging functional properties of groups of genes in the system, but they do not represent or infer causal relationships (Emmert-Streib, Glazko et al. 2012). Likewise, Ingenuity Mechanistic Networks are useful exploratory tools, but are hypothesis networks defined by the knowledge base and number of input genes from a data set. They also have a tendency to 'over-fit' networks, for example TGF-β1 was defined as the upstream regulator of over 700 genes in one analysis of Affymetrix expression data (section **7.3.3**). In addition to being 'scale-free' (Barabási and Oltvai 2004) gene regulatory networks are parsimonious in design, i.e. sparse (Leclerc 2008). This was an emerging issue in **Chapter 5** where a complex and heterogenous network was developed and was insufficiently sparse. There are a number of available strategies for inferring regulatory networks from gene expression surveys. These include correlation-based methods (WGCNA), mutual information (ARACNE) and Boolean

and Bayesian network methods (Liu 2015). Future work would focus on inferring regulatory networks from gene expression data arising from these studies.

Data, data everywhere

Data constraints were a core issue in this thesis. Issues of cost, platform obsolescence, quality control and missing data all arose with microarray gene expression profiling studies in this thesis. To overcome limitations in the data arising from novel experimental work, and to place gene expression profiles in a wider context, methods of gene expression data integration were considered by data merging approaches. Many practical concerns were evident including the poor quality of many expression profiles submitted to public repositories, difficulty in handling data from multiple microarray platforms, the dearth of available cartilage and tendon profiles, the minimal sequence and target overlap between probes from different platform technologies and the problems associated with handling noisy and heterogenous data.

Although some guidelines for microarray meta-analysis and data integration have been suggested (Ramasamy, Mondry et al. 2008) there is no standard procedure for dealing with diverse expression data. Diversity of the data is a function of the plethora of commercial products and lack of standardization across the musculoskeletal research community with respect to disease models and laboratory methods. Although the methodologies used in this thesis require refinement, in terms of inclusion criteria and the focusing of research questions, they represent the first attempt to describe regulatory sub-networks associated with native cartilage and tendon as well as three-dimensional culture systems.

8.5: Project objectives redefined

An understanding of dedifferentiation as a regenerative mechanism must develop with respect to other regenerative models, i.e. stem cells (Maden 2013). A body of evidence defines the presence of tissue-resident stem cells in the tendon (Bi, Ehirchiou et al. 2007, Lui and Chan 2011) and chondro-progenitors in the superficial zone of articular cartilage (Candela, Yasuhara et al. 2014). Numerous studies have considered the directed differentiation of MSCs (often from bone-marrow or adipose tissue) towards a chondrocytic phenotype (Boeuf and Richter 2010) with view to autologous sources of stem cells for cartilage repair. The findings of developmental marker expression in monolayer cells in this study would have benefitted from analysis in a wider context, for example, transcriptomic profiling alongside differentiating MSCs or comparisons relative to the monolayer culture methodologies used for deriving tendon-derived stem cells. This would facilitate a narrower definition of dedifferentiation by delineating the extent of any developmental phenotype. Elucidation of a restricted de-/re-differentiation regulatory network would be a valuable contribution to the modeling of novel organotypic culture systems and informing regenerative musculoskeletal interventions.

References

Barabási, A.-L. and Z. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nature Reviews: Genetics **5**(2): 101-113.

Bi, Y., D. Ehirchiou, T. Kilts, C. Inkson, M. Embree, W. Sonoyama, L. Li, A. Leet, B.-M. Seo, L. Zhang, S. Shi and M. Young (2007). "Identification of tendon stem/progenitor cells and the role of the extracellular matrix in their niche." Nature Medicine **13**(10): 1219-1227.

Boeuf, S. and W. Richter (2010). "Chondrogenesis of mesenchymal stem cells: role of tissue source and inducing factors." Stem Cell Research & Therapy **1**(4): 31.

Boogerd, F., F. Bruggeman and R. Richardson (2013). "Mechanistic Explanations and Models in Molecular Systems Biology." Foundations of Science **18**(4): 725-744.

Candela, M. E., R. Yasuhara, M. Iwamoto and M. Enomoto-Iwamoto (2014). "Resident mesenchymal progenitors of articular cartilage." Matrix Biology **39**(0): 44-49.

Chikina, M., E. Zaslavsky and S. Sealfon (2015). "CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations." Bioinformatics (Oxford, England) **31**(10): 1584-1591.

De Ceuninck, F., C. Lesur, P. Pastoureau, A. Caliez and M. Sabatini (2004). Culture of Chondrocytes in Alginate Beads. Cartilage and Osteoarthritis, Vol. 1: Cellular and Molecular Tools. M. Sabatini, P. Pastoureau and F. De Ceuninck, Humana Press Inc. **100:** 15.

Deshmukh, A., M. Murgia, N. Nagaraj, J. Treebak, J. Cox and M. Mann (2015). "Deep Proteomics of Mouse Skeletal Muscle Enables Quantitation of Protein Isoforms, Metabolic Pathways, and Transcription Factors." Molecular & Cellular Proteomics **14**(4): 841-853.

El-Gohary, Y., S. Tulachan, J. Wiersch, P. Guo, C. Welsh, K. Prasadan, J. Paredes, C. Shiota, X. Xiao, Y. Wada, M. Diaz and G. Gittes (2014). "A smad signaling network regulates islet cell proliferation." Diabetes **63**(1): 224-236.

Emmert-Streib, F., G. Glazko, G. Altay and R. de Matos Simoes (2012). "Statistical inference and reverse engineering of gene regulatory networks from observational expression data." Frontiers in Genetics **3**: 8.

Gaujoux, R. and C. Seoighe (2013). "CellMix: a comprehensive toolbox for gene expression deconvolution." Bioinformatics (Oxford, England) **29**(17): 2211-2212.

Gilad, Y. and O. Mizrahi-Man (2015). "A reanalysis of mouse ENCODE comparative gene expression data. ." F1000Research **4**: 121. Under peer-review.

Haycock, J. (2011). 3D Cell Culture: A Review of Current Approaches and Techniques. 3D Cell Culture. J. Haycock, Humana Press. **695:** 1-15.

Hsueh, M.-F., P. Önnerfjord and V. B. Kraus (2014). "Biomarkers and proteomic analysis of osteoarthritis." Matrix Biology **39**: 56-66.

Jiang, Y. and R. S. Tuan (2015). "Origin and function of cartilage stem/progenitor cells in osteoarthritis." Nat Rev Rheumatol **11**(4): 206-212.

Jopling, C., S. Boue and J. C. Izpisua Belmonte (2011). "Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration." Nature Reviews. Molecular Cell Biology **12**(2): 79-89.

Kapacee, Z., S. Richardson, Y. Lu, T. Starborg, D. Holmes, K. Baar and K. Kadler (2008). "Tension is required for fibripositor formation." Matrix Biology **27**(4): 371-375.

Kragl, M., D. Knapp, E. Nacu, S. Khattak, M. Maden, H. H. Epperlein and E. Tanaka (2009). "Cells keep a memory of their tissue origin during axolotl limb regeneration." Nature **460**(7251): 60-65.

Krämer, A., J. Green, J. Pollard and S. Tugendreich (2014). "Causal analysis approaches in Ingenuity Pathway Analysis." Bioinformatics **30**(4): 523-530.

Leclerc, R. (2008). "Survival of the sparsest: robust gene networks are parsimonious." Molecular Systems Biology **4**(1).

Lin, S., Y. Lin, J. Nery, M. Urich, A. Breschi, C. Davis, A. Dobin, C. Zaleski, M. Beer, W. Chapman, T. Gingeras, J. Ecker and M. Snyder (2014). "Comparison of the

transcriptional landscapes between human and mouse tissues." Proceedings of the National Academy of Sciences of the United States of America **111**(48): 17224-17229.

Liu, Z.-P. (2015). "Reverse Engineering of Genome-wide Gene Regulatory Networks from Gene Expression Data." Current Genomics **16**(1): 3-22.

Lui, P. P. Y. and K. M. Chan (2011). "Tendon-Derived Stem Cells (TDSCs): From Basic Science to Potential Roles in Tendon Pathology and Tissue Engineering Applications." Stem Cell Reviews **7**(4): 883-897.

Maden, M. (2013). "Who needs stem cells if you can dedifferentiate?" Cell Stem Cell **13**(6): 640-641.

Nehrt, N., W. Clark, P. Radivojac and M. Hahn (2011). "Testing the ortholog conjecture with comparative functional genomic data from mammals." PLoS Computational Biology **7**(6): e1002073.

Porrello, E., A. Mahmoud, E. Simpson, J. Hill, J. Richardson, E. Olson and H. Sadek (2011). "Transient regenerative potential of the neonatal mouse heart." Science (New York, N.Y.) **331**(6020): 1078-1080.

Prasad, A., S. Kumar, C. Dessimoz, S. Bleuler, O. Laule, T. Hruz, W. Gruissem and P. Zimmermann (2013). "Global regulatory architecture of human, mouse and rat tissue transcriptomes." BMC Genomics **14**(1): 716.

Prestwich, G. (2007). "Simplifying the extracellular matrix for 3-D cell culture and tissue engineering: a pragmatic approach." Journal of Cellular Biochemistry **101**(6): 1370-1383.

Prockop, D. (2009). "Repair of Tissues by Adult Stem/Progenitor Cells (MSCs): Controversies, Myths, and Changing Paradigms." Molecular Therapy **17**(6): 939-946.

Ramasamy, A., A. Mondry, C. Holmes and D. Altman (2008). "Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets." PLoS Med **5**(9): e184.

Sass, S., F. Buettner, N. Mueller and F. Theis (2014). "RAMONA: a Web application for gene set analysis on multilevel omics data." Bioinformatics (Oxford, England).

Schulze-Tanzil, G. (2009). "Activation and dedifferentiation of chondrocytes: Implications in cartilage injury and repair." Annals of Anatomy - Anatomischer Anzeiger **191**(4): 325-338.

Seok, J., S. Warren, A. Cuenca, M. Mindrinos, H. Baker, W. Xu, D. Richards, G. McDonald-Smith, H. Gao, L. Hennessy, C. Finnerty, C. López, S. Honari, E. Moore, J. Minei, J. Cuschieri, P. Bankey, J. Johnson, J. Sperry, A. Nathens, T. Billiar, M. West, M. Jeschke, M. Klein, R. Gamelli, N. Gibran, B. Brownstein, C. Miller-Graziano, S. Calvano, P. Mason, P. Cobb, L. Rahme, S. Lowry, R. Maier, L. Moldawer, D. Herndon, R. Davis, W. Xiao and R. Tompkins (2013). "Genomic responses in mouse models poorly mimic human inflammatory diseases." Proceedings of the National Academy of Sciences of the United States of America **110**(9): 3507-3512.

Shay, T., J. Lederer and C. Benoist (2015). "Genomic responses to inflammation in mouse models mimic humans: we concur, apples to oranges comparisons won't do." Proceedings of the National Academy of Sciences of the United States of America **112**(4).

Spanoudes, K., D. Gaspar, A. Pandit and D. Zeugolis (2014). "The biophysical, biochemical, and biological toolbox for tenogenic phenotype maintenance in vitro." Trends in Biotechnology **32**(9): 474-482.

Sugimoto, Y., A. Takimoto, H. Akiyama, R. Kist, G. Scherer, T. Nakamura, Y. Hiraki and C. Shukunami (2013). "Scx+/Sox9+ progenitors contribute to the establishment of the junction between cartilage and tendon/ligament." Development (Cambridge, England) **140**(11): 2280-2288.

Szibor, M., J. Pöling, H. Warnecke, T. Kubin and T. Braun (2014). "Remodeling and dedifferentiation of adult cardiomyocytes during disease and regeneration." Cellular and Molecular Life Sciences : CMLS **71**(10): 1907-1916.

Takao, K. and T. Miyakawa (2015). "Genomic responses in mouse models greatly mimic human inflammatory diseases." Proceedings of the National Academy of Sciences of the United States of America **112**(4): 1167-1172.

Thomas, P., V. Wood, C. Mungall, S. Lewis and J. Blake (2012). "On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report." PLoS Computational Biology **8**(2): e1002386.

Worthley, Daniel L., M. Churchill, Jocelyn T. Compton, Y. Tailor, M. Rao, Y. Si, D. Levin, Matthew G. Schwartz, A. Uygur, Y. Hayakawa, S. Gross, Bernhard W. Renz, W. Setlik, Ashley N. Martinez, X. Chen, S. Nizami, Heon G. Lee, H. P. Kang, J.-M. Caldwell, S. Asfaha, C. B. Westphalen, T. Graham, G. Jin, K. Nagar, H. Wang, Mazen A. Kheirbek, A. Kolhe, J. Carpenter, M. Glaire, A. Nair, S. Renders, N. Manieri, S. Muthupalani, James G. Fox, M. Reichert, Andrew S. Giraud, Robert F. Schwabe, J.-P. Pradere, K. Walton, A. Prakash, D. Gumucio, Anil K. Rustgi, Thaddeus S. Stappenbeck, Richard A. Friedman, Michael D. Gershon, P. Sims, T. Grikscheit, Francis Y. Lee, G. Karsenty, S. Mukherjee and Timothy C. Wang (2015). "Gremlin 1 Identifies a Skeletal Stem Cell with Bone, Cartilage, and Reticular Stromal Potential." Cell **160**(1–2): 269-284.

Yao, L., C. S. Bestwick, L. A. Bestwick, N. Maffulli and R. M. Aspden (2006). "Phenotypic drift in human tenocyte culture." Tissue Engineering **12**(7): 1843-1849.

Young, A., M. Smith, S. Smith, M. Cake, P. Ghosh, R. Read, J. Melrose, D. Sonnabend, P. Roughley and C. Little (2005). "Regional assessment of articular cartilage gene expression and small proteoglycan metabolism in an animal model of osteoarthritis." Arthritis Research & Therapy **7**(4): R852-861.

Zi, Z., D. A. Chapnick and X. Liu (2012). "Dynamics of TGF-β/Smad signaling." FEBS Letters **586**(14): 1921-1928.

[END]